Sluttrapport for pilotering av KDRS produksjonslinje hos Giske kommune og politiet

Sammendrag

Denne rapporten beskriver pilotering av KDRS produksjonslinje hos Giske kommune og politiet. Det gjøres også en vurdering av KDRS sin produksjonslinjemetodikk.
Pilotene har avdekket både styrker og svakheter ved produksjonslinjemetoden. Funnene viser at metoden er svært gunstig for arkivskaper, både med tanke på tidsbruk og kostnader. Dette understøtter antagelsen om at metoden vil bidra til at arkivmateriale som har gått ut av administrativ bruk, raskere overføres til depot og at etterslepet vil bli redusert. Samtidig fører metoden til at en del arbeidsoppgaver overføres fra arkivskaper til depot. Det gjelder særlig arbeidet med å beskrive og behandle arkivmaterialet. Gjenbruk av malsett gjør imidlertid at denne belastningen er overkommelig og forskjellen i praksis er ikke nødvendigvis så stor.
Prosjektets vurdering av metoden opp mot de arkivfaglige hensynene autentisitet, integritet, pålitelighet og anvendbarhet viser at de tre første later til å være tilstrekkelig ivaretatt ved bruk av metoden. Det er en risiko knyttet til anvendbarhet, dersom en ikke får samlet tilstrekkelig og riktig informasjon om originalsystemet. I slike tilfeller kan det bli vanskelig, i verste fall umulig, å tolke hele eller deler av innholdet i SIARD-databasen på sikt. Arkivverket og KDRS er oppmerksomme på denne risikoen og vil utforme retningslinjer for hvordan arkivmateriale skal beskrives, hvilken tilleggsinformasjon som er viktig og hvordan en bør gå frem for å få tak i denne tilleggsinformasjonen.
Rapporten konkluderer med at produksjonslinjemetoden er egnet for å ta inn etterslepet av arkivmateriale, med de forutsetningene som er beskrevet i avsnittet om bruksområdet for metoden. Det anbefales derfor at Riksarkivaren gir godkjenning for bruk av metoden på Noark-systemer for kommunal sektor. For øvrige systemer kreves ingen godkjenning.

Bakgrunn

KDRS produksjonslinje for bevaring og formidling av elektroniske arkiv fra kommunal sektor (videre omtalt som KDRS produksjonslinje) ble utarbeidet med hovedmål om å finne en brukervennlig produksjonslinje som produserer kvalitativt gode arkivpakker. Riksrevisjonens rapporter fra 2010 og 2017 peker på store mengder digitalt skapte arkiver i kommunal og statlig sektor som står i fare for å gå tapt. Disse arkivene skal oppbevares i kortere eller lengre tid og defineres i det følgende som etterslepet. MAVOD-rapporten[1] anslo at etterslepet i kommunal sektor utgjør uttrekk fra ca. 2200 systemer og med dagens metoder vil det kreve over 500 årsverk å få dette til depot. SAMDOK-rapporten[2] anslo i 2014 at 17 fylkeskommuner i snitt har 25 systemer, totalt 475 systemforekomster, mens 428 kommuner i snitt her 28,5 systemer totalt 12192 systemforekomster. Tallene er presisert som et minimumstall. Rapporten anslår at 60-65% av systemforekomstene er bevaringsverdige. Med bakgrunn i dette så KDRS-miljøet et behov for å finne en ny metode for å hente inn arkivene. Denne metoden skulle bli et supplement til Noark-uttrekk og tabelluttrekk beskrevet med ADDML. De ønsket at den nye metoden skulle være enkel i bruk for å senke terskelen og gjøre det enklere å fremstille arkivuttrekk. Dette var bakgrunnen for at KDRS utviklet en egen produksjonslinje.

I denne rapporten beskriver og vurderer vi KDRS sin produksjonslinjemetodikk. Det er gjennomført to piloter, en i kommunal sektor og en i statlig sektor.

Den kommunale piloten er gjennomført i samarbeid mellom Arkivverket, KDRS, IKA Møre og Romsdal og Giske kommune. Beskrivelsene er i stor grad utformet basert på innspill fra KDRS-miljøet og vurderinger er gjort av ansatte i Arkivverket.

Den statlige piloten er gjennomført med politiet. Den har vært mer omfattende for Arkivverket, hovedsakelig fordi Arkivverket her har fungert som depot.

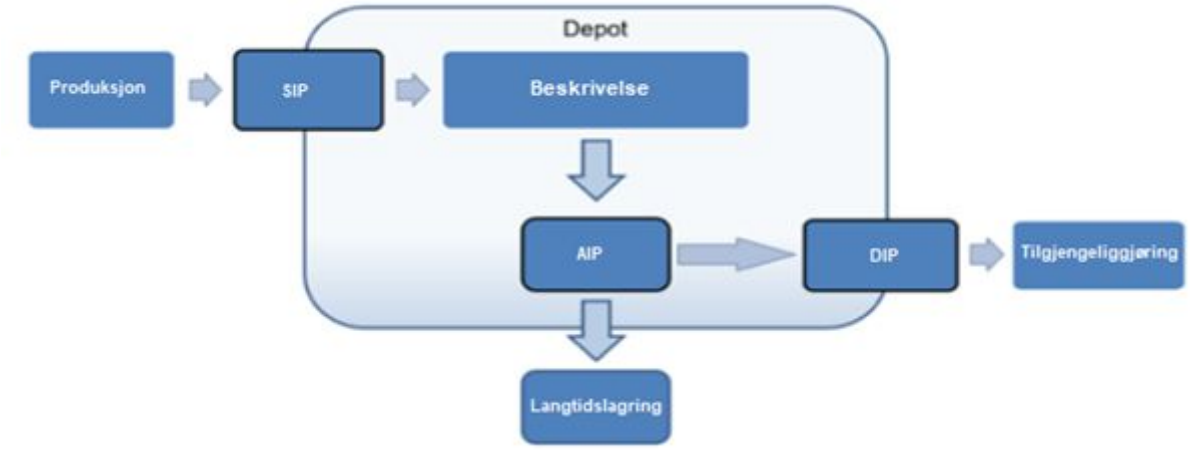
En detaljert beskrivelse av det som har blitt gjort finnes i vedlegg 1 og 2, som beskriver fremgangsmåten for pilotene.

[1] MAVOD rapport: https://www.arkivverket.no/arkivutvikling/utviklingsprosjekter/mavod
[2] SAMDOK rapport 2014-2: Digitalt skapt materiale i kommunal sektor 1985-2010 kartlegging https://samdokdotcom.files.wordpress.com/2015/01/rapport-samdok-2014_2-kartlegging-av-digitalt-skapt-materiale-i-kommunal-sektor.pdf

Beskrivelse av KDRS Produksjonslinje for bevaring og formidling av elektroniske arkiv fra kommunal sektor

I avsnittene under beskriver vi hva som skjer i de ulike trinnene i produksjonslinjemetodikken. Håndtering av mottatt digitalt arkivmateriale i depot ble utført i henhold til "KDRS prosesser og rutiner" [3]. Disse rutinene ligger til grunn for prosesstrinnene som inngår i produksjonslinjen.

Prosessen gjelder både for Noark-uttrekk og tabelluttrekk, men i det følgende fokuseres det på tabelluttrekk i SIARD-formatet. Figuren under viser hovedtrinnene i prosessen. Disse beskrives i nærmere detalj i de neste avsnittene.



Produksjonslinjemetoden er basert på OAIS-modellen og er i henhold til DIAS-pakkestruktur[4].

Produksjon

For å imøtekomme behovet om at det skal være så enkelt som mulig for arkivskaper, besluttet KDRS at arkivskaper skulle fremstille en SIARD-versjon av databasen. Dette er en kostnadseffektiv måte å sikre informasjon fra relasjonsdatabaser og kan gjøres av arkivskaper uten involvering av en tredjepart.

SIARD

SIARD-formatet ble utviklet av det sveitsiske riksarkivet i 2007 og oppdateres kontinuerlig. Hensikten med formatet var at det skulle kunne brukes til å arkivere relasjonsdatabaser og være leverandøruavhengig. En SIARD-fil er en ZIP-fil som inneholder

[3] http://www.kdrs.no/prosjekt/ProsessDigitaltDepot.htm
[4] https://www.arkivverket.no/forvaltning-og-utvikling/regelverk-og-standarder/dias-prosjektet-digital-arkivpakkestruktur?q=dias

- En speiling av innholdet i databasen
- Maskinprosesserbare strukturelle metadata

SIARD-filen kan også inneholde tekstfiler og binære filer (CLOBer og BLOBer) internt, eller referere til eksterne filer. Det betyr at formatet er fleksibelt med hensyn til opprinnelig systems måte å lagre dokumenter. Formatet bygger på standardene XML og SQL:2008. Gjeldende versjon er SIARD 2.1.

Det er mulig å se på de ulike tabellene i et SIARD-uttrekk uten spesialverktøy.
SIARD-databaser lar seg migrere tilbake til en relasjonsdatabase for å gjøre endringer.
Views er en del av SIARD-standarden og SIARD-verktøy kan dokumentere SQL-spørringer fra de opprinnelige views. Disse lagres i SIARD metadata.xml.
Disse kan videre brukes som grunnlag for en innsynsversjon av SIARD-uttrekket, enten i en SIARD-viewer, migrert tilbake til en annen database (eksempelvis MySQL) eller transformert til et format som støttes av en innsynsløsning.

De strukturelle metadataene er lagret i filen metadata.xml. Filen er en eksakt teknisk beskrivelse av relasjonsdatabasens struktur, det vil si databaser, skjema, tabeller, felter, felttyper, relasjoner og i noen tilfeller også views og SQL-spørringer. For hver tabell beskrives plassering i den hierarkiske mappestrukturen, et fritekstfelt, en liste over kolonner (inkludert navn og datatype), antall rader, identifikasjon av primærnøkler og andre tekniske detaljer. Filen metadata.xml kan også inneholde strukturelle detaljer som views, routines, constraints og triggers. Mer teknisk informasjon om SIARD-formatet kan finnes hos eksempelvis Library of Congress[5] .

SIARD-uttrekksverktøy

Det finnes flere ulike verktøy som kan benyttes for å genere SIARD-uttrekk. For eksempel Spectral Core Full Convert (SCFC), Siard Suite, Database Preservation Toolkit (DPTK) m. fl. SCFC er lisensbelagt programvare, mens de to andre er open source.

KDRS produksjonslinje støtter bruk av alle SIARD-verktøyene, men erfaringen til KDRS er at SCFC fungerer på flere databaser enn de andre programvarene. SCFC støtter et stort antall databaseplatformer og har aktiv brukersupport. Derfor anbefaler IKA Møre og Romsdal sine kommuner å bruke dette verktøyet for å generere SIARD-uttrekkene og av denne grunn er det verktøyet som har blitt benyttet i disse pilotene.

[5] https://www.loc.gov/preservation/digital/formats/fdd/fdd000426.shtml

Beskrivelse

Decom

Depot mottar SIARD-databasen fra arkivskaper og importerer den inn i Decom-programvaren. Decom er lisensbelagt programvare utviklet av Documaster[6].
Her beskrives SIARD-databasen ved å tilføre metadata i form av tabell-, felt og nøkkelbeskrivelser. Det er også mulig å prioritere de ulike tabellene ut i fra hvor viktige de er, samt å markere systemtabeller som ikke inneholder informasjon. Prioritering av tabeller brukes i test, validering og kvalitetssikring i depot og det er tenkt at dette også skal brukes til senere visning. Decom lagrer disse metadataene i et malsett som man kan dele. Dette innebærer at det er nok at kun én bruker av Decom skriver inn alle metadataene, mens alle som senere skal beskrive samme system gjenbruker de samme metadataene via malsettet. De som gjenbruker malsettet må selv sjekke det opp mot systemet de bevarer og eventuelt supplere med metadata som ikke allerede er beskrevet. Denne type gjenbruk er antatt å være særlig relevant for kommunal sektor som har mange forekomster av samme system og effektiviserer dermed jobben i depot. Decom konverterer også dokumenter til arkivformatet PDF/A via Microsoft Word eller LibreOffice, etter å ha analysert hvilke filtyper som finnes i uttrekket. Decom endrer også filendelsen for dokumenter som allerede er i arkivformat, eksempelvis TIFF, JPEG eller TXT. Output fra Decom er en SIARD-fil med ekstra metadata og konverterte arkivdokumenter (eksternt eller internt i SIARD-filen), samt den orginale SIARD-filen og malsettet i JSON-format.

Malsett

SIARD-databasens metadata.xml blir i Decom transformert til Decom JSON malsett. Et Decom JSON malsett er en kopi av de elementer som er nødvendig for å kunne gjenbruke et malsett for å beskrive uttrekk av samme type system. Det inneholder den eksakte tekniske strukturen av databasesystemets skjema, tabeller, felter, felttyper og relasjoner som SIARD-standardens metadata.xml inneholder. I tillegg suppleres hver enkelt tabell, felt og relasjon med beskrivelser via malsettet. KDRS har begynt å lage retningslinjer for hvordan man kan bruke entiteter som viser hvordan de ulike tabellene relaterer seg til Noark 4 - og 5-format. Malsettbeskrivelsene gjenbrukes når malsettet brukes på et annet uttrekk av samme system og det er her effektiviseringsgevinsten ligger. For systemer med flere hundre tabeller, med mange felter i hver tabell vil det være svært tidsbesparende å kun være nødt til å kvalitetssikre at beskrivelsene på tabellene og feltene.

Den tekniske beskrivelsen av en relasjonsdatabase, de strukturelle metadata, er identiske om formen på disse er som Decom JSON malsett, SIARD metadata.xml eller ADDML. Transformasjon mellom Decom JSON malsett, SIARD metadata.xml og (den tekniske metadata-delen av) ADDML er mulig og en teoretisk mapping er spesifisert av KDRS.

[6] https://www.documaster.com/no/documaster_decom

Selve Decom-programvaren eies av Documaster, mens malsettene er fristilt fra denne og eies av KDRS / offentlig sektor.

Utvikling av malsett

For at produksjonslinjemetoden skal være mest mulig effektiv er det viktig at malsettene som blir utviklet har tilstrekkelig kvalitet. I 2018 - 2019 har KDRS i samarbeid med Arkivverket via MODARK-prosjektet utført et dugnadsprosjekt for å utvikle malsett for eldre systemer med høyest forekomst i kommunal sektor. Målsetningen er å utvikle og kvalitetssikre malsett tilhørende systemene som vises i tabellen under:

Navn	Leverandør	Kategori
DocuLive	SI (Tieto)	1-01 Arkivtjeneste-saksarkiv
ephorte	Evry	1-01 Arkivtjeneste-saksarkiv
ESA	Evry	1-01 Arkivtjeneste-saksarkiv
Forum Winsak98	Ergo (Evry)	1-01 Arkivtjeneste-saksarkiv
Kontor 2000	EDB (Evry)	1-01 Arkivtjeneste-saksarkiv
Sofie	Unique (Visma)	1-01 Arkivtjeneste-saksarkiv
Symfoni	Cinet	1-01 Arkivtjeneste-saksarkiv
Gaia HMS	HMSsystemer	1-03 Sentrale-systemer
Oppvekst Barnehage	Visma	4-06 Barnehage
Extens	IST	4-09 Skole
Sats Skole	IST	4-09 Skole
WIS Skole	Waade Information System	4-09 Skole
PPI	Unique (Visma)	4-10 PPT
PPT-HK	HK data	4-10 PPT
BVPro	Hiadata (Visma)	6-14 Barnevern
Familia	Visma	6-14 Barnevern
Marthe	Unique (Visma)	6-14 Barnevern
Vaktdata	Visma	6-14 Barnevern
MD Flyktning	Visma	6-15 Flyktningetjeneste
Winmed Helse	Profdoc	6-16 Helsetjeneste
Gerica	Tieto	6-17 Pleie-og-omsorg
Profil	Visma	6-17 Pleie-og-omsorg
Acos Sosial	Dips	6-18 Sosialtjeneste
Oskar	Unique (Visma)	6-18 Sosialtjeneste
Velferd	Visma	6-18 Sosialtjeneste

Flere kommunearkivinstitusjoner har bidratt i denne utviklingen. Malsettene er utviklet med støtte fra arkivutviklingsmidlene og vil derfor være tilgjengelig også for arkivinstitusjoner som ikke er medlem av KDRS. Tilgangen til malene må imidlertid styres for å ivareta leverandørenes interesser. Til dette formålet har KDRS utviklet et eget avtaleverk.

Dokumentasjon fra det opprinnelige system og bruken av dette

Metadatabeskrivelsen fra Decom gir informasjon om struktur og innholdselement og er essensiell for å kunne tolke dataene i ettertid. Men for å kunne gjenbruke og tilgjengeliggjøre SIARD-databasen er det behov for mer informasjon enn det man får tilført gjennom Decom og malsettene. Sammenlignet med et Noark-uttrekk hvor informasjonsmodellen for arkivuttrekket er kjent er informasjonsmodellen for SIARD-databasen i utgangspunktet ukjent. Det vil si at kunnskapen om relasjonene mellom tabellene, koblinger, nøkler med mer som skjer i applikasjonslogikken i produksjonssystemet er ukjent. Dersom denne informasjonen mangler kan det bli utfordrende å gjenbruke og tolke dataene i ettertid. For å løse dette gjøres det derfor grundige undersøkelser for å finne og tolke de nødvendige sammenhengene. I noen tilfeller opprettes det også støttedokument med viktig informasjon om produksjonssystemet. Eksempelvis for Doculive i politipiloten ble det for viktige felter i malsettet definert opp noter [NOTE XXX] som ble utdypet i et tekstdokument. Det inneholdt blant annet informasjon om filstier. Depot samarbeider også med arkivskaper for å kvalitetssikre det de har kommet frem til og de vurderingene som er gjort, samt for å hente inn supplerende informasjon. Depot samler inn det som finnes av brukerdokumentasjon, manualer og lignende, samt tar screenshots fra systemet av de skjermbildene som har blitt identifisert som relevante. I tillegg intervjues arkivskaper for å avdekke bruksmønster av systemet, for eksempel om det er brukt på en annen måte enn det som fremgår av systemdokumentasjonen. Det er startet et arbeid i Arkivverket for å se på hvordan dette kan inngå i en arkivbeskrivelse.

Testing og kvalitetssikring

For å sikre at arkivversjonen holder tilstrekkelig kvalitet blir det utført en rekke tester. Depot utfører blant annet DROID-analyser[7] av arkivdokumentene før og etter at Decom har konvertert dem og sammenligner resultatene, for å sikre at konverteringen var vellykket. I tillegg kjøres det ulike PDF-valideringer for å sikre at de konverterte dokumentene er i PDF/A-format, samt kontroll av filstiene. Metoden har per i dag ikke etablert et testregime for mottakskontroll, dette er noe man må se på og sektoren må enes om.

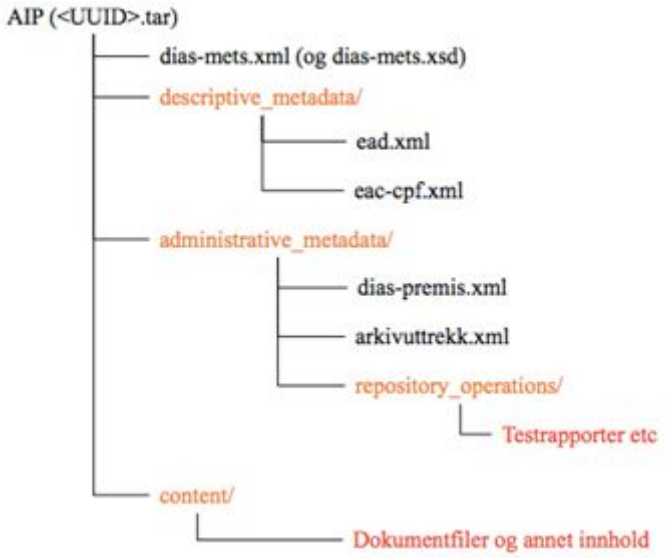
På bakgrunn av den innsamlede informasjonen fremstiller depot også en enkel formidlingsvariant som kvalitetssikres med arkivskaper. Hensikten med dette er å sikre at arkivversjonen gjenspeiler det opprinnelige systemet og at informasjonen er den samme. Mer informasjon om hvordan dette ble gjort i pilotene finnes i vedleggene.

---

7 http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/

Langtidslagring

Informasjonen fra beskrivelsen inngår deretter i en AIP som er klar for å legges i et sikringsmagasin. Denne AIPen er i henhold til DIAS-pakkestruktur, illustrert i figuren under.



SIARD-filen og arkivdokumentene legges under content-mappen. Rapporter fra DROID, PDF-validering, Arkade 5, antivirussjekk mv. plasseres under repository_operations under administrative_metadata-mappen. Her plasseres også forretningslogikk, views og annen dokumentasjon man har samlet sammen fra det opprinnelige systemet. Det er foreløpig ikke gitt bestemte retningslinjer for hvordan dette skal struktureres.

Opprettelse av arkivuttrekk.xml for et fagsystem er ikke en etablert praksis i produksjonslinjen. Hovedgrunnen til dette er fordi spesifikasjonen for teknisk beskrivelse ADDML 8.3 av XML-uttrekk ikke finnes. Den tekniske metadatainformasjonen fra ADDML-beskrivelsen finnes i SIARDs metadata.xml.

Utover dette suppleres AIPen med metadata på samme måte som ethvert annet arkivuttrekk.

Tilgjengeliggjøring

Produksjonslinjemetodikken omfatter også tilgjengeliggjøring. Per i dag er imidlertid ikke de tekniske løsningene som trengs for tilgjengeliggjøring på plass. Når disse løsningene kommer på plass er det nødvendig å lage DIPer av de AIPene som skal gjøres tilgjengelig.

Prosjektets vurdering

Vi vil i de neste avsnittene vurdere metoden og funnene fra pilotene opp mot arkivfaglige egenskaper, regelverk, ressursbruk og andre praktiske hensyn.

Vurdering av metoden opp mot arkivfaglige egenskaper

Tilliten til arkiv er avhengig av tilstrekkelig ivaretakelse av hensynene autentisitet, anvendbarhet, pålitelighet og integritet. Dette er beskrevet i ISO-standard 15489, om informasjon, dokumentasjon og dokumentasjonsforvaltning. I de følgende avsnittene beskriver vi disse egenskapene med utgangspunkt i standarden, samt gjør en vurdering av produksjonslinjemetodikken sett opp mot hver av dem.

Anvendbarhet

En forutsetning for at vi kan ha tillit til arkiv er at arkivmaterialet er anvendbart og tilgjengelig for bruk. I dette ligger det at materialet kan gjenfinnes, hentes frem, presenteres og tolkes. I ettertid bør det kunne presenteres i direkte forbindelse til forretningsaktiviteten eller transaksjonen som gav opphav til den. Ved vurdering av anvendbarhet er det nyttig å ta utgangspunkt i fire nivåer:

1. Materialet finnes. Dette innebærer at materialet ikke er tapt.
2. Materialet er teknisk tilgjengelig. Dette innebærer at materialet kan fremstilles for videre bruk.
3. Materialet er forståelig. Dette innebærer at materialet kan tolkes.
4. Materialet kan gjenbrukes. Dette innebærer at materialet kan benyttes videre av andre virksomheter og i andre sammenhenger.

Produksjonslinjemetodikken sikrer at materialet finnes og sørger for at det ikke går tapt. Ved å fremstille en SIARD-versjon av produksjonsdatabasen får depot all informasjonen som fantes i produksjonssystemet, forutsatt at en tar med alle tabeller i SIARD-versjonen. Vi anser at materialet er teknisk tilgjengelig ettersom det er gjort systemuavhengig og kan lastes inn i en annen relasjonsdatabase.

For å sikre at materialet er forståelig er det i tillegg nødvendig at både databasestrukturen og den overliggende informasjonsmodellen og applikasjonslogikk som er relevant for tolkning er beskrevet. Disse beskrivelsene følger ikke med automatisk når SIARD-filen opprettes. En viktig del av produksjonslinjemetodikken er derfor å opprette malsett som inneholder beskrivelser av databasestrukturen. Som en del av dette undersøkes databasen for å finne og tolke de nødvendige sammenhengene. I tillegg tas screenshots av relevante skjermbilder, det samles inn brukerdokumentasjon og i de tilfeller det er tilgjengelig, hentes views i form av SQL-spørringer ut av systemet. Deretter kvalitetssikres disse beskrivelsene med arkivskaper. I en del tilfeller vil databasestrukturen se annerledes ut enn den overliggende informasjonsmodellen, særlig kan dette tenkes å være tilfelle i mer komplekse

systemer. I slike tilfeller trengs også informasjonsmodellen for å fullt ut kunne forstå materialet. Beskrivelse av den overliggende informasjonsmodellen er per i dag ikke en del av av produksjonslinjemetodikken. For å fullt ut kunne tolke materialet for ettertiden er det også behov for gode arkiv- og aktørbeskrivelser av arkivet.

For å sikre at viktig materiale kan tolkes anbefaler prosjektet at det etableres detaljerte retningslinjer for denne delen av produksjonslinjen.

Vi legger til grunn at det er nødvendig å kjenne til de vesentligste delene av informasjonsmodellen for at materialet skal kunne gjøres tilgjengelig på en ressurseffektiv måte. Både KDRS og Arkivverket jobber for tiden med å finne ut hvordan digitalt skapt materiale kan gjøres tilgjengelig og det er foreløpig ikke avklart nøyaktig hvilke krav som må stilles til arkivmaterialet for dette formålet. Det som imidlertid virker klart er at AIPen som kommer ut av produksjonslinjen krever mer arbeid før den kan tilgjengeliggjøres. Funnene fra denne piloten tyder på at informasjonen som trengs for å gjøre materialet tilgjengelig ligger i AIPen, blant annet slik at en informasjonsmodell kan beskrives i ettertid.

Integritet

Denne egenskapen handler om hvorvidt arkivmaterialet er fullstendig og uendret. I dette ligger det en vurdering om hvorvidt arkivmaterialet har vært gjenstand for uautorisert endring. For å kunne ha tillit til arkiv er det nødvendig at autoriserte merknader, tillegg eller slettinger er eksplisitt angitt og er sporbare. For å kunne vurdere om arkivmaterialets integritet er ivaretatt må de som skal forvalte og bruke materialet ha innsikt i hva som har skjedd med arkivmaterialet etter at det ble tatt ut fra systemet der det ble til. Vurderingen henger sammen med graden av transformasjon og hvorvidt prosessene arkivmaterialet har vært igjennom er transparente og veldokumenterte eller ikke.

Det er først og fremst arkivdanningsfasen som legger grunnlaget for materialets integritet. Ved fremstilling av arkivversjoner er det således mer et spørsmål om den integriteten materialet allerede har, opprettholdes eller svekkes. I produksjonslinjemetodikken påføres det sjekksummer på SIARD-databasen for å sørge for at integriteten opprettholdes. Produksjonslinjen baserer seg for øvrig på OAIS-rammeverket og DIAS-pakkestruktur. Arkivpakken lagres i et forvaltningssystem på samme måte som andre pakker. Integriteten er derfor godt ivaretatt. En fordel med denne metoden er at den er transparent og at det ikke gjøres noen endringer på dataene underveis i prosessen. I tillegg har man alltid mulighet til å gå tilbake til databasen slik den opprinnelig var, gitt at arkivpakken som ble sendt inn (IP0) blir tatt vare på.

Autentisitet

Med autentisitet menes at arkivmaterialet er hva det hevder å være, at materialet har blitt produsert eller sendt av den personen eller instansen som hevder å ha produsert eller sendt det og videre at det er produsert eller sendt på det påståtte tidspunktet. I ISO-standarden er autentisiteten knyttet til metadata som påføres arkivmaterialet. Arkivskapere må ha rutiner

på plass som sikrer at slike metadata påføres. Dette bør skje tidlig i arkivdanningen, så nærme transaksjonen som mulig.

I likhet med integritet, legges grunnlaget for autentisitet tidligere i verdikjeden. Produksjonslinjemetodikken vil derfor ikke ha noen direkte innvirkning på dette. De metadata som er påført i arkivdanningen blir bevart når arkivpakken lages med produksjonslinjemetodikken og autentisiteten opprettholdes dermed.

Pålitelighet

Arkivmateriale som er pålitelig har et innhold som gir en fullstendig og nøyaktig gjengivelse av de transaksjoner, aktiviteter og fakta som det dokumenterer, og som vi dermed kan stole på og bruke som grunnlag i videre transaksjoner og aktiviteter. Påliteligheten er avhengig av at metadata som dokumenterer materialets kontekst påføres så tett opp til transaksjonen som mulig, både i tid og sted. Påføringen av slike metadata må settes opp eller gjøres av personer som har direkte kunnskap om hvordan transaksjonen har funnet sted i virksomheten.

Som ved integritet og autentisitet er vurderingen av pålitelighet knyttet til hvorvidt de metadata som allerede er påført bevares i produksjonslinjemetodikken. De metadata som er påført i arkivdanningen blir bevart når arkivpakken lages med produksjonslinjemetodikken og påliteligheten opprettholdes dermed.

Vurdering av metoden opp mot relevant regelverk og standarder innen arkivområdet

Dette avsnittet inneholder en vurdering av produksjonslinjemetoden opp mot arkivloven, arkivforskriften, riksarkivarens forskrift, personvernforordningen og personopplysningsloven.

Arkivloven, arkivforskriften, riksarkivarens forskrift og ISO-standarder

Arkivloven inneholder overordnede bestemmelser om arkiv og disse bestemmelsene utdypes i arkivforskriften og riksarkivarens forskrift. Det overordnede kravet som følger av arkivloven § 6 er at dokumentene skal sikres som informasjonskilder for samtid og ettertid. Alle andre bestemmelser i arkivloven handler dypest sett om å ivareta dette. Arkivforskriften gir utdypende regler og riksarkivarens forskrift gir mange og detaljerte regler og stiller tekniske krav om hvordan arkivskaperne skal handle for å oppfylle arkivlovens og arkivforskriftens krav. Det er særlig i riksarkivarens forskrift vi treffer bestemmelser som har direkte betydning for produksjonslinjemetoden. Dette gjelder særlig bestemmelsene om periodisering (kapittel 4), formater (kapittel 5) og bevaring og kassasjon (kapittel 7). Samtidig er det prosjektets vurdering at det kan være fullt mulig å oppnå en kvalitetsmessig forsvarlig bevaring i tråd med både arkivloven og arkivforskriften selv om man skulle tillate at arkivskaperne går frem på den måten, og ved hjelp av de standardene produksjonslinjemetoden legger opp til. Dette kan utløse behov for å vurdere unntak fra enkelte forskriftsbestemmelser eller egne forskriftsbestemmelser for å ta igjen etterslepet.

Denne rapporten skal gi Riksarkivaren en del av grunnlaget for å beslutte om produskjonslinjemetoden bør tas i bruk og hvilke tiltak som er hensiktsmessige.

I det følgende vurderer vi de bestemmelsene i disse kapitlene som er relevante for produksjonslinjemetoden.

Innenfor ISO-området finnes det flere standarder som er relevante for arkiv. Disse standardene har ikke status som hverken lov eller forskrift, og ligger dermed «under» disse. De er imidlertid nyttige og gir støtte i å vurdere hvordan man kan ivareta sentrale arkivfaglige begreper som autentisitet, integritet, pålitelighet og anvendbarhet.

Periodisering

Kapittel 4 i riksarkivarens forskrift inneholder bestemmelser om periodisering. Det er særlig § 4-6 om klargjøring av elektronisk journal og arkiv for deponering i arkivdepot som er relevant. Det vesentlige her er å se på formålet med periodisering. Det handler om å legge til rette for en hensiktsmessig avgrensning av innholdet i et arkiv, som kan bevares som en avgrenset helhet, slik at man kan ta uttrekk av denne del av arkivet, basert på perioden. Av denne bestemmelsen følger det at en elektronisk kopi av journal- og arkivsystemet for den avsluttede arkivperioden skal klargjøres for deponering i arkivdepot. Ettersom SIARD-databasen er en eksakt kopi av produksjonsdatabasen er det i praksis ikke mulig å implementere periodisering på det materialet som ligger i produksjonsdatabasen. Den mest realistiske implementeringen av periodisering er å bruke skarpt periodeskille på det tidspunktet man tar SIARD-uttrekket. De øvrige bestemmelsene om periodisering er ikke relevante for produksjonslinjemetoden, ettersom disse blir ivaretatt av sakarkiv- eller fagsystemet, før uttrekkstidspunktet.

I en database med flere arkivdeler/-perioder skal man kunne ta uttrekk av enkelte arkivdeler, uten å måtte ta hele databasen. Dersom man tar flere SIARD-uttrekk fra samme database risikerer man å ende opp med ulike versjoner av et dokument, dersom det har blitt gjort endringer i produksjonssystemet. Dette kan skape usikkerhet rundt hvilket dokument som er det originale. Denne problemstillingen er ikke unik for denne metoden og SIARD-formatet og er en del av en større diskusjon.

Dokumentformater

Kapittel fem i riksarkivarens forskrift er omfattende og inneholder detaljerte bestemmelser om lagringsmedier, format og avleveringer for elektronisk materiale. For kommuner og fylkeskommuner gjelder kun de bestemmelsene i kapittelet slik det er fastsatt i kapittel tre og åtte. Når det gjelder produksjonslinjemetoden er det kun reglene som er fastsatt gjennom kapittel tre som er relevante, ettersom kapittel åtte retter seg mot konvertering i ettertid for digital bevaring. I § 5-5 går det fram at kapitelet først og fremst retter seg mot statlige arkivskapere som avleverer til Arkivverket, mens for kommuner og fylkeskommuner gjelder reglene i kapitel 5 bare i den grad det går fram av kapitel 3. Dermed gjelder bestemmelsene i kapittel 5 §§ 5-17 til 5-20 for kommuner og fylkeskommuner, i og med at det vises til disse det vises til i kapitel 3, jf § 3-3 nr 1.

Bestemmelsene i §5-17 - §5-20 retter seg mot arkivdokumenter og handler om krav til tekniske filformater, pakking og komprimering av data mv. Produksjonslinjemetoden gjør det mulig å håndtere enkeltfiler, slik at arkivdokumenter kan konverteres mellom ulike filformater. Dette kan også gjøres batchvis, slik at store mengder arkivdokumenter kan konverteres som en operasjon. Tekniske krav til arkivdokumenter kan derfor imøtekommes slik en ellers ville gjort i andre uttrekksmetoder, uten at det påvirkes av verken SIARD-formatet eller produksjonslinjemetoden.

Særlig om produksjonslinjemetoden og Noark-system

Riksarkivarens forskrift §§ 5-21 til 5-23 inneholder krav til I henhold til riksarkivarens forskrift § 3-3 tredje ledd er kommunal sektor pålagt å følge de bestemmelsene i forskriftens kapittel fem som gjelder fremstilling av arkivversjoner fra Noark-systemer. Det gjelder blant annet bestemmelsene §§ 5-21 og 5-23. I § 3-3 fjerde ledd i samme forskrift åpnes det for andre eksport- og dokumentformater for kommunal sektor, forutsatt at disse godkjennes av Riksarkivaren. Så lenge en slik godkjenning ikke foreligger kan ikke produksjonslinjemetoden benyttes som eneste metode for å ta uttrekk fra Noark-system. Arkivverket og KDRS er i dialog om en godkjenning på dette punktet, og denne rapportens anbefaling er at Riksarkivaren gir en slik godkjenning, se forøvrig konklusjoner og anbefalinger.

Krav til teknisk og administrativ dokumentasjon

Del VI beskriver nærmere krav til teknisk dokumentasjon og metadata for arkivuttrekket, blant annet at den tekniske dokumentasjonen skal følge reglene i Riksarkivarens beskrivelsesstandard ADDML. Produksjonslinjemetoden er ikke i tråd med regelverket på dette punktet. SIARD metadata.xml ivaretar imidlertid hensynet til den delen av ADDML som omhandler strukturelle metadata. Dersom andre deler av beskrivelsesstandarden skal ivaretas må disse metadataene suppleres på annet vis. Det er mulig å konvertere SIARD metadata.xml til ADDML og supplere med mer informasjon her. Det arbeides for tiden med en oppgave som omhandler metadata i arkiv, og fremtidig praksis bør være i tråd med utfallet fra dette arbeidet.

Produksjonslinjen er videre i tråd med de øvrige krav til teknisk og administrativ dokumentasjon, samt krav til lagringsmedium og filorganisering som beskrevet i lovverkets §5-26 - §5-32, da dette dekkes av DIAS-standardene og depots øvrige prosedyrer, blant annet bruk av Arkade 5.

Bevaring og kassasjon

Kapittel 7 inneholder bestemmelser om bevaring og kassasjon og består i hovedsak av henholdsvis generelle bevarings- og kassasjonsbestemmelser for egenforvaltningssaker i statlige organer (del II) og generelle bevarings- og kassasjonsbestemmelser for fylkeskommunale og kommunale arkiv skapt etter 1950 (del III). For produksjonslinjemetoden brukt i kommunal sektor er det del III som er relevant. Hensikten med bestemmelsene er primært å sikre at bevaringsverdig arkiv blir tatt vare på for ettertiden.

Produksjonslinjemetoden tar i utgangspunktet vare på hele databasen. Selv om slik merbevaring utover det som er pålagt bevart i Riksarkivarens forskrift kapitel 7, i utgangspunktet ikke er i strid med gjeldende arkivregler, jf § 7-23 som gjelder for kommunal sektor, forutsetter reglene at det gjøres en reell vurdering av bevaringsverdi i tilknytning til slik merbevaring. I tillegg vil personvernforordningen og personopplysningslovens krav føre til at vurderinger av bevaringsverdi og nødvendighet er påkrevd.

For at innholdet skal ha verdi som arkiv er det imidlertid en forutsetning at det beskrives slik at det kan forstås. I praksis er det derfor beskrivelsen av malsettet, og gjennom det beskrivelsen av databasen, som implementerer bevaringsbestemmelsene i §7-24 - §7-33. Det er derfor viktig at den eller de som beskriver malsettet er godt kjent med bevaringsbestemmelsene vist til over og er i stand til å identifisere og beskrive denne informasjon. Gitt at dette imøtekommes vil produksjonslinjemetoden kunne imøtekomme disse kravene. Tydelige retningslinjer for hvordan malsett skal beskrives, og at dette gjøres som et teamarbeid, vil lette dette arbeidet.

Bevaringsbestemmelsene nevnt over er minimumskrav. I bestemmelsene §7-22 og §7-23 omtales kassasjon og merbevaring. Utover bevaringsbestemmelsene kan kommune selv vurdere kassasjon og eventuelt merbevaring. Kommunene står fritt til å bevare mer enn det som fremgår av minimumsbestemmelsene. Produksjonslinjemetoden har per i dag ikke et prosesstrinn for kassasjon, utover det som gjøres i systemet før uttrekkstidspunktet. Denne problemstillingen er ikke unik for denne metoden. Dette er imidlertid ikke et problem opp mot bestemmelsene i arkivforskriften, jf. over, men kan være problematisk fra et personvernståsted. Dette er beskrevet nærmere under.

Personvernforordningen og personopplysningsloven

Personvernforordningen gjelder som norsk lov, og personopplysningsloven utfyller denne på enkelte punkter der det er adgang til å gi bestemmelser i nasjonal lovgivning Begge regelverk inneholder en rekke bestemmelser som er relevante for behandlingen av personopplysninger kommunale og interkommunale arkivinstitusjoner gjennomfører som del av arbeidet med å langtidsbevare arkiv.

Produksjonslinjemetoden fører ikke til noen praktisk endring i dagens situasjon med hensyn til behandling av personopplysninger hos de kommunale og interkommunale arkivinstitusjonene. Også med produkslinjemetoden vil det være en utfordring for de kommunene som i dag gjennomfører arkivuttrekk uten å ta stilling til kassasjon, og dermed merbevarer uten noen reell vurdering av bevaringsverdien av materialet, vil kunne få utfordringer med å oppfylle personvernkrav dersom man tar SIARD uttrekk av hele systemer. I verste fall kan kommunen da stå uten lovlig behandlingsgrunnlag for deler av materialet, og brudd på et slikt helt grunnleggende krav til personvern i henhold til personvernforordningen artikkel 6 vil være alvorlig. I tillegg vil flere av de grunnleggende personvernprinsippene i artikkel 5 gjør det tydelig at det er nødvendig å ta aktivt stilling til kassasjon i tilknytning til et arkivuttrekk, jf prinsippene om formålsbegrensning, dataminimering og lagringsbegrensning (artikkel 5 nr 1 bokstav b), c) og e)).

Det kan dermed innvendes at fraværet av muligheten for kassasjon på uttrekkstidspunktet ved bruk av produksjonslinjemetoden er problematisk med hensyn til personvernkrav ved at det potensielt følger med flere personopplysninger i uttrekket enn det som er nødvendig. Ettersom kassasjon av slike opplysninger som ikke er bevaringsverdig skal foretas før uttrekket genereres, kan dette gjøres ved hjelp av systemet som ble brukt for å behandle opplysningene.

Det er ikke noe i metoden som forhindrer kassering underveis i prosessen og dersom noen krever sletting eller minimering av personvernhensyn etter at materialet er deponert kan depot alltid gå inn i SIARD-filen og slette de relevante dataene. Det vil også være mulig å gjøre endringer i ettertid. På dette punktet skiller ikke metoden seg fra dagens metoder.

Opphavsrett

Malsettene kan sees på som en beriket versjon av den eksakte tekniske strukturen av databasesystemets skjema, tabeller, felter, felttyper og relasjoner. Det er uavklart hvorvidt disse databasestrukturene som malsettene bygger på er opphavsrettslig vernet. Vi legger imidlertid til grunn at databasestrukturen er en del av rettighetene og konkurransefortrinnene til den enkelte leverandør, selv om den enkelte kunde, kommune i denne sammenheng, eier dataene som ligger i databasestrukturen.

Som nevnt over er en av de største gevinstene ved å bruke malsett at de kan gjenbrukes på tvers av offentlige virksomheter. Dette forutsetter at malsettene deles på tvers av offentlige virksomheter. For å sikre at dette skjer på en måte som ivaretar leverandørenes kommersielle interesser, og hindrer andre kommersielle aktører innsyn i databasestrukturene, har KDRS etablert et avtaleverk. Avtaleverket består av tre deler:

1. Avtale om forvaltning av rettigheter knyttet til malsett for databasebeskrivelse
2. Bruksavtale for malsett
3. Brukervilkår for KDRS produksjonslinjen

Del 1 av avtaleverket regulerer forholdet mellom den enkelte leverandør og KDRS som forvalter av malsettene. Del 2 regulerer forholdet mellom KDRS og den enkelte offentlige virksomhet som benytter malsettene. Del 3 er brukervilkårene som den enkelte ansatte aksepterer ved å ta i bruk malsettene.

Offentlig sektor er avhengig av et godt samarbeid med leverandørene for at arbeidet med bevaring av arkiv skal kunne skje så enkelt og effektivt som mulig. KDRS tar derfor systematisk kontakt med alle leverandører som eier systemer det lages malsett av, med sikte på å inngå avtaler som beskrevet over.

KDRS har fått avtaleverket kvalitetssikret av advokat, med hensyn til opphavsrett.

Ressursbruk

Arkivskaper

Arbeidet med å fremstille en SIARD-versjon av produksjonssystemet er enkelt og svært tidseffektivt og krever lav grad av IT-kompetanse. Den som skal utføre jobben hos arkivskaper må ha de riktige tilgangene internt for å kunne hente ut dataene med et SIARD-verktøy. Arkivskaper involveres igjen mot slutten av prosessen for kvalitetssikring av arkivversjonen. For Giske kommunes del tok jobben med å beskrive, generere og pakke SIARD-filen, inkludert overføring til depot tok 1 dag og 3 timer.

At metoden er enkel og effektiv å bruke for arkivskaper senker terskelen for å overføre arkivmateriale til depot og øker sannsynligheten for at informasjonen sikres for langtidsbevaring.

Den største økonomiske besparelsen ser ut til å være for systemer som ved dagens metoder krever involvering av en tredjepart. I slike tilfeller vil det være et innsparingspotensiale ved å benytte denne metoden. Det er i tillegg nærliggende å anta at tidsbruken for arkivskapers del blir lavere, da det ikke er behov for oppfølging og koordinering av en tredjepart.

Arkivdepot

Sammenlignet med metodikk for tabelluttrekk, som beskrives i lovverket, får depotinstitusjonen med produksjonslinjemetoden ansvar for en større del av prosessen ettersom de skal beskrive arkivversjonene og konvertere tilhørende dokumenter. Begge disse oppgavene gjøres i utgangspunktet i Decom-programvaren. Per i dag må Decom-arbeidet i mange tilfeller suppleres med en del manuelle operasjoner og skripting. KDRS jobber mot å få implementert mest mulig av dette inn i Decom og ønsker å få dette automatisert så langt det er mulig.

Det er ikke nødvendigvis en ulempe at arkivdepot håndterer disse oppgavene. Eksempelvis vil man på denne måten ha mulighet til å dokumentere hvilke verktøy som har blitt brukt, eventuelle avvik og detaljer rundt konverteringen. Dette kan være nyttig for senere bruk av materialet. Selv om lovverket i dag tilsier at arkivskaper skal produsere fullverdige arkivversjoner viser erfaring fra Arkivverket at mye av arbeidet med å produsere ADDML-filer også gjøres av depot i dag. Det er derfor ikke nødvendigvis slik at dette vil utgjøre en betydelig økning i arbeidsmengden for depot.

I depot effektiviseres jobben med å beskrive uttrekket ved å gjenbruke malsettene. Tidsbruken avhenger av størrelsen og kompleksiteten til system som beskrives. I pilotene har denne delen av prosessen vært nokså tidkrevende, i hovedsak fordi det har blitt jobbet med å kvalitetssikre og forbedre Decom-programvaren parallelt. Etterhvert som verktøyene blir bedre er det forventet at dette vil ta kortere tid. Det anslås at den faktiske tidsbruken i piloten for å importere til Decom, tilpasse malsettet, kontrollere innhold og fremstille en Decom-arkivversjon tok totalt 3 dagsverk for depot i piloten med Giske kommune. I

politipiloten ble det brukt omlag 4 dagsverk. Dette inkluderte da også prototypetesting av formidlingsversjon.

Til sammenligning anslås det at samme beskrivelse for Familia med ADDML ville ta omlag 20 dager, basert på tidsbruken det tok å lage malsettet for Famillia i Decom. Selve malsettet som ble tilpasset var utviklet før piloten startet av medlemmer i KDRS. Dette arbeidet tok 22 dager.

Metoden krever at arkivdepot har løst lisens for Decom-programvaren. Her påløper årlige kostnader. Den enkelte kommune trenger ikke å løse lisens.

Utvikling av malsett

Ressursbruken for å utvikle et malsett avhenger av størrelse og kompleksitet på originaldatabasen. Basert på statistikk fra KDRS har det hittil blitt brukt mellom 2 og 22 dager å utvikle malsettene. Gjennomsnittstiden for utvikling av et malsett er i underkant av 9 dager. Kvaliteten på malsettene er forventet å bli gradvis bedre etter hvert som de brukes og stadig justeres og oppdateres som en del av dette.

Vurdering av metoden opp mot praktiske hensyn

Fremstilling av arkivversjon

Produksjonslinjemetoden er under utvikling og er på noen områder fremdeles umoden. Det er identifisert flere trinn i metoden som har forbedringspotensiale, særlig med tanke på Decom-programvaren og utvikling av malsettene. En svakhet med metoden er at den har mange manuelle arbeidsoperasjoner og ad-hoc-løsninger for å bøte på mangler. Metoden krever i dag gode database- og programmeringsferdigheter. I tillegg må en kjenne til mangler og svakheter i Decom og hvordan det skal kompenseres for dette. KDRS-miljøet har løsninger for de fleste av disse problemene. Det er utviklet en mengde script, både av KDRS og Arkivverket, for å håndtere og tilpasse system med spesielle implementasjoner, eksempelvis databaser med oppstykkede filer, databaser med dokumenter på innsiden / utsiden av SIARD-filen mv. Det planlegges å få mer automatiserte løsninger for dette på sikt, men per i dag er dette noe som må håndteres manuelt. I mellomtiden bør disse verktøyene deles i en verktøybank eller lignende, sammen med retningslinjer for beste praksis.
I nyeste versjon av Decom (v.1.1.3) har KDRS-miljøet nylig fremstilt arkivversjoner fra systemene PPI og Familia uten behov for manuelle justeringer. Det er forventet at mange av punktene i avsnittet over vil forbedres i tiden fremover.

Bevaring

Bevaringsoperasjoner i depot, som mediemigrering, formatkonvertering og overvåking av bitråte, blir ikke påvirket av at arkivversjonen inneholder en SIARD-fil. Gjenfinning av arkivpakker vil heller ikke kreve tilpasninger i depot, ettersom arkivpakkene som lages med produksjonslinjemetoden følger DIAS-standarden for pakkestruktur og er søkbare på metadata på samme måte som andre arkivpakker.

Tilgjengeliggjøring

Produksjonslinjen legger opp til at tilgjengeliggjøring skal inngå i metoden, men per i dag eksisterer det ikke en løsning for dette. Det pågår egne prosjekter både hos KDRS og Arkivverket som ser på tilgjengeliggjøring.

Selv om detaljene rundt tilgjengeliggjøring er uklare, vurderer prosjektet det slik at AIPene som lages er mulig å bearbeide for tilgjengeliggjøring, forutsatt at beskrivelsene er gode nok. Aktuelle format på en visningspakke (DIP) kan eksempelvis være SIARD-format, Noark 4/5 uttrekksformat, et generisk tabelluttrekk iht. en generisk standard for publisering eller noe annet.

Piloten med politiet har vist at en beriket SIARD-fil i seg selv kan være kilden til en visningspakke som vises i en innsynsmodul. Uavhengig av om det bygges videre på denne prototypen eller om det besluttes å følge et annet spor, bør en slik innsynsmodul ha egne views for innsyn tilpasset brukernes behov. På sikt kan man se for seg informasjon om slike view kan bli en del av et Decom-malsett. I KDRS-miljøet har Noark 5-struktur hittil vært

vurdert som den mest sannsynlige strukturen for visningspakker levert gjennom Decom. Dette fordi man da kan bruke eksisterende KDRS-innsynløsning for Noark 5-uttrekk, eventuelt importere Noark 5-uttrekk inn i en standard Noark 5-kjerne og bruke visningsmodulene der.

KDRS jobber med å implementere bruk av maskinlesbare entiteter i malsettene som utvikles. Dette vil gi større frihet rundt valg av DIP-format og muligheter for eventuell omstrukturering til annen informasjonsmodell på et senere tidspunkt, særlig med hensyn til gjenfinning og tilgjengeliggjøring. I fremtiden kan det bli aktuelt å legge inn dataanalyse, automatisk tagging, beskrivelse og tester som en del av produksjonslinjen. Deler av malsettet kan leses maskinelt på den måten det er lagt inn i dag.

I alle tilfeller blir videreutvikling og standardisering av beskrivelser og entiteter i malsettene viktig. Det samme gjelder forvaltning av navnekonvensjoner. Bruken av disse bør være del av en brukerveiledning. Denne type skriveregler, for å bøte på mangler i løsningene, innfører et nytt lag av kompleksitet og er ikke en ideell løsning.

Basert på de overordnede undersøkelsene som er utført i piloten mener prosjektet det er grunnlag for å anta at det vil være mulig å tilgjengeliggjøre innholdet i arkivversjonene som fremstilles ved hjelp av produksjonslinjemetoden.

Sårbarhet knyttet til programvare

Slik produksjonslinjen er satt opp i dag har man mange valg når det gjelder verktøy for å generere SIARD-databasen, men kun ett verktøy (Decom) for å beskrive databasen. Denne mangelen av alternativer for verktøystøtte for beskrivelse av databaser representerer en mulig sårbarhet ved metoden. Denne sårbarheten er blant annet relevant fordi Documaster er et relativt lite og ungt selskap.

Selv om Decom er lisensiert programvare er formatet for malsettene åpent. Det innebærer at andre leverandører også kan utvikle programvare for beskrivelse og gjenbruk av malsett. Dette er positivt og gir mulighet for større valgfrihet for verktøystøtte på sikt. I tillegg er kunnskapen som ligger til grunn for metoden i KDRS-miljøet, og ikke låst til en leverandør.

Ettersom det eksisterer flere SIARD-uttrekksverktøy er det mindre sårbarhet knyttet til prosesstrinnet som omhandler fremstilling av en SIARD-versjon. Her finnes det også verktøy som er open source.

Sårbarhet knyttet til forvaltning

Mye av kunnskapen om metoden er i dag knyttet til et begrenset antall enkeltpersoner. Dette representerer en sårbarhet. Det er dialog mellom KS, KDRS og Arkivverket om hvordan metoden og malsettene kan forvaltes på nasjonalt nivå, på en måte som gjør både metode, malsett og verktøy tilgjengelig for alle kommuner og fylkeskommuner.

Selve SIARD-formatet forvaltes i dag av det Sveitsiske Riksarkivet og er i bruk i flere land. Vi anser det som lite sannsynlig at formatet skal opphøre eller ikke forvaltes videre.

Bruksområdet for produksjonslinjemetoden

Basert på funnene i pilotene er vår vurdering at metoden vil være egnet for å hente inn etterslepet av eldre system i kommunal sektor, ettersom de fleste av disse systemene baserer seg på enkeltdatabaser og kompleksiteten i databasene er overkommelig. I tillegg går de samme systemene igjen på tvers av kommunene og malsettene vil derfor kunne gjenbrukes, med mindre justeringer. I prinsippet vil også metoden kunne brukes på sakarkivsystemer der uttrekksfunksjonaliteten ikke fungerer som den skal. Det kan for eksempel være tilfeller hvor det er nødvendig med mye tilpasning eller ordning i databasen for at uttrekket skal kunne produseres med akseptabel kvalitet, for eksempel at de fleste journalposter innen perioden(e) kommer med. Det betyr at det vil dreie seg om tilfeller hvor arkivskaper (m.fl.) har prøvd å produsere ordinære uttrekk, og det kan påvises at det ikke har latt seg gjøre å få til akseptable uttrekk. Ettersom denne rapporten omhandler etterslepet vil dette stort sett være snakk om klassiske sakarkiv-systemer, hovedsakelig Noark 4. Noark 5-sytemer skal i utgangspunktet ikke være så gamle at de faller inn under denne kategorien. I motsetning til uttrekk fra øvrige system vil imidlertid dette kreve en godkjenning av Riksarkivaren, jf. riksarkivarens forskrift § 3-3. Arkivverket jobber for tiden med dette i samarbeid med KDRS.

Metoden baserer seg på SIARD-formatet og er derfor i utgangspunktet avhengig av at arkivmaterialet som skal bevares er lagret i en relasjonsdatabase. Ettersom metoden tar utgangspunkt i relasjonsdatabaser er det også enkelte andre forhold som har betydning for når den er egnet:

- Metoden er meget godt egnet der løsningen benytter én relasjonsdatabase, hvor det i liten grad brukes applikasjonslogikk til å fremstille informasjonen i databasen og der hvor kompleksiteten i databasen ikke er for høy.
- Metoden vil gi størst effekt der systemet det gjelder benyttes mange steder, slik at malsettet kan gjenbrukes.
- Det er uvisst om metoden vil være egnet for systemer der arkivmaterialet er spredd over flere databaser, dette må undersøkes nærmere.
- Metoden er egnet for komplekse systemer med høy forekomst i sektoren hvor det benyttes mye applikasjonslogikk, men jobben med å lage malsett blir større.
- Metoden vil være mindre egnet for databaser som er spesielt komplekse og inneholder et stort antall tabeller, med få forekomster i sektoren. Jobben med å lage malsett blir særlig omfattende og gjenbruksgevinsten lav.

Det er mer tvilsomt om produksjonslinjemetoden og bruk av SIARD-databaser er egnet for fremtidige systemer, jf. kulepunktene over. Det skyldes blant annet at det stadig oftere benyttes andre typer databaser enn relasjonsdatabaser og at kompleksiteten i løsningene øker. I slike tilfeller bør en vurdere andre alternativer for å sikre dokumentasjonen, for eksempel gjennom forhåndsdefinerte grensesnitt. Dette forutsetter imidlertid at man har tatt

hensyn til overføring av arkiv allerede da systemet ble spesifisert og at grensesnittet allerede finnes i systemet.

Konklusjoner og anbefalinger

Piloten har avdekket både styrker og svakheter ved produksjonslinjemetoden. Funnene viser at metoden er svært gunstig for arkivskaper, både med tanke på tidsbruk og kostnader. Dette understøtter antagelsen om at metoden vil bidra til at arkivmateriale som har gått ut av administrativ bruk, raskere overføres til depot og at etterslepet vil bli redusert. Samtidig fører metoden til at en del arbeidsoppgaver overføres fra arkivskaper til depot. Det gjelder særlig arbeidet med å bearbeide og beskrive arkivmaterialet. Gjenbruk av malsett og script gjør imidlertid at denne belastningen er overkommelig. I tillegg har kommunale arkivinstitusjoner en tradisjon for å være aktive overfor arkivskaperne de betjener og har også før produksjonslinjemetoden ble tatt i bruk, bistått i beskrivelse av arkiv. Forskjellen i praksis er derfor ikke nødvendigvis så stor.

Vår vurdering av metoden opp mot de arkivfaglige hensynene autentisitet, integritet, pålitelighet og anvendbarhet viser at de tre første later til å være tilstrekkelig ivaretatt ved bruk av metoden. Det er en risiko knyttet til anvendbarhet, dersom en ikke får samlet tilstrekkelig og riktig informasjon om originalsystemet. I slike tilfeller kan det bli vanskelig, i verste fall umulig, å tolke hele eller deler av innholdet i SIARD-databasen på sikt. Denne risikoen forsterkes av at det foreløpig ikke er etablert retningslinjer for hvordan denne informasjonen skal samles inn. Både Arkivverket og KDRS er oppmerksomme på denne risikoen og vil utforme retningslinjer for hvordan arkivmateriale skal beskrives, hvilken tilleggsinformasjon som er viktig og hvordan en bør gå frem for å få tak i denne tilleggsinformasjonen. I tillegg arbeides det med en innsynsløsning for arkivpakker som er laget ved hjelp av produksjonslinjemetoden.

Det er en utfordring ved metoden at den foreløpig stiller forholdsvis høye kompetansekrav til de som skal benytte den, blant annet innen database- og programmeringsferdigheter. Dette fører både til en sårbarhet ved at man blir avhengig av et lite antall ressurspersoner, samtidig som det skaper utfordringer med å oppnå utbredelse av metoden. For å avhjelpe dette videreutvikles Decom-programvaren, slik at den blir mer brukervennlig og det skal lages brukerveiledninger for metoden.

Et viktig premiss fra MAVOD-rapporten er at en for arkivmateriale som står i fare for å gå tapt kan godta lavere kvalitet for å sikre at materialet kommer inn til depot, selv om det vil kunne medføre økt ressursbehov i depot. En må vurdere det faglig ideelle opp mot hva som er mulig å få til innenfor de rammene som gjelder. Det gjelder blant annet kompetanse, ressurser og økonomi hos arkivskaper.

På denne bakgrunn vurderer vi at produksjonslinjemetoden er egnet for å ta inn etterslepet av arkivmateriale, med de forutsetningene som er beskrevet i avsnittet om bruksområdet for metoden. Det anbefales derfor at Riksarkivaren gir godkjenning for bruk av metoden på Noark-systemer for kommunal sektor. For øvrige systemer kreves ingen godkjenning.

Vedlegg 1: Giske kommune

Dette vedlegget beskriver fremgangsmåten for piloten med Giske kommune.

For å teste produksjonslinjen tok vi utgangspunkt i Familia.

Forarbeid Arkivskaper

Det ble laget et uttrekk fra en Oracledatabase med SCFC som resulterte i en fil på SIARD 2.1-format. Denne jobben ble utført av IT-personell hos eKommune Sunnmøre. Jobben med å generere SIARD-uttrekket tok omlag 1 time. Det gikk også med 1 time til pakking av uttrekket med Arkade 5, samt 1 time å kryptere og oversende.

Arbeid i Depot

SIARD-uttrekket ble lastet inn i Decom v.1.1.3, hvor det ble arbeidet i flymodus uten nettilgang på en kryptert disk. De utarbeidede malsettene ble lastet ned i forkant, og malsettet for Familia ble valgt. Alle tabellene ble gått gjennom og det ble gjort en grovsortering. Det ble sjekket match mot alle tabellene i malsettet, men en gikk ikke inn på de enkelte feltene. Tabellene ble prioritert som høy, medium eller lav (eller “no priority” for tomme tabeller). Det ble ikke gjort endringer i tekstbeskrivelsene i malsettet.

Det ble utført en Droid-analyse på den opprinnelige SIARD-databasen, for å sjekke filstruktur og for å få oversikt over hvilke filtyper som inngikk i SIARD-versjonen. Det ble sjekket hvor mange ulike filtyper det finnes og antall forekomster av hver type og det ble generert en rapport. BLOBer ble deretter konvertert i Decom, via LibreOffice. Dette skjedde parallelt med kvalitetskontrollen av malsettet. Resultatet fra Decom ble sammenlignet og verifisert med DROID-analysen. Det ble utført to ulike PDF-valideringer på de konverterte dokumentene ved bruk av VERA PDF. 37757 filer ble konvertert til gyldig PDF/A ihht analysene. 61 binærfiler og et fåtall andre filer kunne ikke konverteres og må sjekkes opp manuelt av depot i samråd med arkivskaper. Dette er ikke blitt gjort som en del av piloten.

Det er anslått at total tidsbruk for depot er 3 dager. Dette inkluderer Decom import, gjenbruk av malsett, kontroll av innhold, fremstilling av arkivversjon og merge til beriket SIARD.

Følgende er ikke med i oppført tid Depot - IKAMR:

- Saksbehandling innmelding og opprette avtale for system
- Tid Decom klient bruker på å lage PDF/A-dokumenteter i DCM (klient kjører uten tilsyn/arbeidstid)
- Lage XQuery, SQL Query o.l. for test og validering av fagsystem tabelluttrekk Familia
- Kjøre XQuery, SQL Query o.l for test og validering av fagsystem tabelluttrekk Familia
- Lage DIP innsynsløsning med URD (beriket SIARD migrert til MySQL med URD innsyn)
- Oppretting av Arkivpakke AIP i ESSArch Tools for Producer

- Opplasting av Arkivpakke AIP til ESSArch Preservation Platform med ESSArch Tools for Archivist

Innhenting av dokumentasjon fra det opprinnelige system og bruken av dette

Videre ble det avholdt et arbeidsmøte for piloten som gikk over to dager, inkludert et møte hos Giske kommune. I løpet av denne tiden ble SIARD-databasen undersøkt for å få mest mulig oversikt over relasjoner. Hensikten med dette arbeidet og møtet med kommunen var å identifisere:

- Databasestrukturen og relasjonene i den
- Viktig informasjon og konverteringer som blir gjort i produksjonssystemet og som ikke kommer med i SIARD-versjonen (applikasjonslogikk)
- Kommunens bruk av systemet. Er det brukt på en måte som det ikke er tiltenkt?

For å finne ut av dette ble databasen migrert over til MySQL og verktøyene phpMyadmin, PostgreSQL og DBeaver ble brukt til å undersøke tabellene. Disse viser full mapping av alle relasjoner i databasen, samt tabeller og felt i et ER-diagram. FA_SAKSJOURNAL og FA_POSTJOURNAL ble ansett som viktige og dermed undersøkt spesielt nøye. Alle tabellene de har relasjoner til ble identifisert. Man prøvde også å finne koblingene mellom tabeller og felter (hvilke er én-til-én, hvilke er mange-til-mange?).

Møte hos Giske kommune

Hos Giske kommune ble det avholdt et møte mellom pilotdeltakerne og en bruker av systemet. Systemet ble undersøkt og man prøvde å nøste opp i sammenhenger. Det er noen tabeller anses som mer interessante enn andre og for disse dokumenteres bruksmønster og det tas relevante skjermbilder. Tabellene Postjournal, Saksjournal, Klientliste, Økonomi og rutiner, Saksbehandling, Rapporten og Journalnotat ble ansett som viktige for Familia. Det ble i møtet gjort noen funn som ikke ville vært lette å komme frem til uten dette besøket. For eksempel kunne man i Journalnotat skrive korte notat i notatfeltet, men dersom de som har behov for å skrive mer utfyllende tekst måtte det genereres en word-fil. Det er dermed inkonsistens mellom feltene i tabellen og relasjoner til dokument. Et annet eksempel er at man finner ut at noen postjournaler ikke er knyttet til klient. Det vil si at det i noen tilfeller ikke finnes klientnummer og dermed mangler dokumentnummer, men det finnes en link til et dokument. Dette er en kobling man må finne ut av i etterkant og som må dokumenteres i arkivversjonen.

Kvalitetssikring av arkivpakken

For å kvalitetssikre arkivpakken ble det laget en enkel visningsvariant av databasen ved hjelp av URD. Hensikten med dette var at brukere fra Giske kommune skulle få kvalitetssikre at søk i arkivpakken ga samme resultater som søk i det aktive Familiasystemet.

Ressursbruk

Jobben med å beskrive, generere og pakke SIARD-filen, inkludert overføring til depot tok 1 dag og 3 timer for Giske Kommune. IKA Møre og Romsdal brukte totalt 3 dagsverk på å ferdigstille en AIP.

Vedlegg 2: Pilot med politiet

Dette vedlegget beskriver piloten med politiet hvor Arkivverket fungerte som depot. Piloten omhandlet uttrekk fra tre politidistrikt, Nordre Buskerud, Søndre Buskerud og Telemark. Testrapporten fra Nordre Buskerud er vedlagt som eget vedlegg for å illustrere resultatet fra testene.

Formål

Formålet med piloten er for Arkivverkets del å skaffe erfaring og empiri rundt produksjonslinjemetoden og verktøyene som brukes. Formålet for politiet er hurtigst mulig deponering av historiske baser, og sanering av gammel teknologi.

Suksesskriterier

Suksesskriteriet for piloten er at den resulterer i godkjente uttrekk.

For å avgjøre dette har vi vurdert fire forhold:

- Ivaretakelsen av de arkivfaglige egenskapene autentisitet, integritet, anvendbarhet og pålitelighet
- Forvaltning i et langtidsperspektiv
- Fremfinning i uttrekkene
- Tilgjengeliggjøring av uttrekkene

Bakgrunn

Politiet var i slutten av 2018 i gang med å anskaffe nytt sakarkivsystem. Som en del av dette ønsket de å deponere 48 historiske databaser fra Doculive-sakarkivsystemet som skulle fases ut. Politiets erfaring fra tidligere deponeringer er at det har vært både tidkrevende og kostbart. De ønsker en enklere prosess for å få overført historisk arkivmateriale til Arkivverket. MODARK-prosjektet hadde behov for å involvere statlige virksomheter i piloteringen av produksjonslinjen. Politiet og deres pågående arbeid med utskifting av sakarkivsystem vil være en god case for slik pilotering.Det ble avtalt at politiet fikk ansvaret for å produsere uttrekket i SIARD-format med utgangspunkt i spesifikasjoner fra Arkivverket. Arkivverket var ansvarlig for å utarbeide malsettet og anvende det på uttrekkene. Politiet ble utover dette involvert ved behov og for avsjekk underveis.

Arkivverket mottok i første omgang SIARD-versjoner av tre ulike systemer fra politiet. Alle disse var basert på Noark 4-baser fra Doculive-systemer og uttrekkene var blitt produsert ved hjelp av verktøyet Spectral Core Full Convert. Systemene inneholdt standard Doculive-tabeller og var i liten grad spesialtilpasset politiet med hensyn til implementering. Da systemene var i drift lå de tilhørende arkivdokumentene i et eksternt fillager med referanser fra Doculive-databasen. Vi mottok en .tar-fil produsert av Arkade 5. SIARD-versjonen lå i content mappen sammen med dokumentene som lå i en egen mappe i content.

Fremgangsmåte

Utvikling av malsett

Det eksisterte ikke et malsett for Doculive og dette måtte derfor utvikles som en del av piloten. Det ble leid inn en ekstern konsulent med ekspertise på Doculive systemer for å sikre at malsettet holdt tilstrekkelig kvalitet. Konsulenten gikk gjennom alle tabeller og førte på beskrivelser på alle tabeller og felt som er viktige for å kunne gjenbruke av databasen. For å forenkle søk og gjenfinning på et senere tidspunkt ble det brukt spesielle skriveregler og tagger på de ulike tabell- og feltbeskrivelsene. Skrivereglene som ble brukt er utviklet av KDRS. Eksempelvis ble tabeller som tilsvarer samme tabell i et Noark 4-uttrekk merket {n4:...} og tilsvarende {n5:...} for Noark 5. Taggene er ihht publiserte Noark-standarder. Tabellene under viser eksempler på tagginger.

Tabell 1: Eksempler på Noark 4-tagginger

{n4:NOARKSAK}	Tabell for Sak
{n4:SA.SAAR}	Felt for Saksår i tabell Sak
{n4:na}	Hvis tabellen i et Noark 4-system ikke kan knyttes Noark 4-spesifikasjonen

Tabell 2: Eksempler på Noark 5-tagginger

{n5:sakaar#saksmappe}	M011 sakaar, her notert som forekomst i Saksmappe
{n5:sakaar}	M011 sakaar, uten at forekomst i Noark 5 v4.0 objekt er angitt
{n5:na}	Hvis tabellen i et Noark 5-system ikke kan knyttes Noark 5-spesifikasjonen

Noen tabeller finnes i både Noark 4- og Noark 5-standarden. I slike tilfeller tas begge feltene med i SIARD-beskrivelsen noe som muliggjør en transformasjon fra SIARD til både Noark 4- og Noark 5-formater, om det skulle bli aktuelt på sikt. Gitt et eksempel med et Noark 4-system hvor tabell for sak, felt for saksår kan uttrykkes som {n4:SA.SAAR}. Samme felt kan også uttrykkes med en Noark 5-tag slik {n5:sakaar#saksmappe}. Begge feltene tas med i SIARD-beskrivelsen og uttrekket får følgende xml-element i metadata.xml beriket:

<description>{n4:SA.SAAR} {n5:sakaar#saksmappe} Saksår</description>

For å beskrive mer avanserte sammenhenger brukes [NOTE-00n], der n er løpenummeret. Eksempelvis er informasjon om fil-lager, fil-informasjon og pekere i Doculive databaser betegnet med [NOTE-002] og satt inn for alle tabellene som inneholder slik informasjon.

Disse sammenhengene er utdypet i et eget dokument, se vedlegg 3. Dokumentet inneholder ellers viktig informasjon om Doculive-malsettet og Doculive-baser generelt.

Enkelte tabell- og feltbeskrivelser er merket med [DIP] via Decom knapp og betyr at innholdet er særlig relevant i en visningsmodul.

Som en del av malsettet ble også de enkelte tabellene prioritert ut i fra hvor viktige de anses å være. Viktigheten kategoriseres som “høy”, “medium”, “lav” eller “no priority”. I tillegg kan man merke systemtabeller som ikke inneholder data med informasjonsverdi som “system”, “dummy”, “statistikk” eller “empty”.

Bearbeidelse av SIARD-filene

SIARD-filene ble importert til Decom hvor de ble matchet med malsettet som ble laget. Det ble gjort en grundig manuell gjennomgang av alle tabeller med prioritet “høy” og “medium”. Det er disse som anses som mest viktige med tanke på gjenbruk av materialet. Her ble alle felter sjekket for å sikre at beskrivelsene fra malsettet var riktig. Dersom et felt har samme navn i malsett og SIARD-filen, samt at beskrivelsen virker riktig ble det antatt at beskrivelsen er riktig. For de resterende tabellene ble kun tabellbeskrivelsene sjekket for å sikre at tabellene inneholder det malsettet tilsier og at prioriteringen ser fornuftig ut.
Arbeidet med å kvalitetssikre malsettet opp mot uttrekket tok omlag 1 time.

Ettersom SIARD-filene vi mottok hadde filreferanser til eksterne dokumenter slik de var i Doculive-basen måtte disse oppdateres. Formålet med dette er å opprettholde lenken mellom eksempelvis journalposter og tilhørende dokumenter, som er essensielt for videre bruk av materialet. På tidspunktet da piloten ble gjennomført hadde ikke Decom støtte for dette og det ble derfor gjort ved hjelp av scripting. For å kunne gjøre denne endringen måtte tabellene og feltene som inneholdt filreferansene identifiseres. Dette ble gjort ved å undersøke databasen, malsettet og det tilhørende dokumentet. Dokumentet beskriver disse sammenhengene godt og viser til hvilke tabeller og felter som inneholder filreferanser.
I piloten tok denne jobben omtrent 8 timer. En av hovedgrunnene til at dette tok tid var fordi vi satt oss inn i fremgangsmåten samtidig som vi prøvde å finne sammenhengene. Dersom man kjenner metoden og vet hvilken informasjon man leter etter, samt har beskrivelsene i støttedokument tilgjengelig, anslås jobben å ta 1-2 timer.

Pronom-filtypeanalyse på filene i SIARD-uttrekket viste at en stor del av filene var i ikke godkjent arkivformat. Decom har støtte, gjennom Libreoffice, for å konvertere filer som ligger inne i selve SIARD-filen. Men siden uttrekket inneholdt filereferanser til filer utenfor dokumentet ble dette gjort med eget script med samme konverteringsbibliotek, samt utvikling av egen metodikk for epost-filer (.msg). For å kunne ettergå scriptet i fremtiden valgte vi å beholde produksjonsformatfilene ved siden av filene som allerede var på arkivformat og de konverterte filene og referere til begge filene i den relevante SIARD-tabellen.
Konverteringsprosessen tok hånd om flesteparten av filene og de filtypene som ikke ble konvertert har enten liten eller ingen bevaringsverdi eller var passordbeskyttet. I tillegg produserte konverteringsjobben sjekksummer som er lagt inn i SIARD-tabellen. Totalt tok konverteringsjobben for 113.000 filer rundt 23 timer, men dette er en passiv prosess og vil

kunne gå ned ved bruk av mer spesialisert programvare og kraftigere maskinvare. Aktiv arbeidstid anslås til under 3 timer.

Samlet anslås det altså at arbeidet med å bearbeide SIARD-filene i depot tok i underkant av ett dagsverk.

Tester for å vurdere metoden

Tester ble definert for å undersøke om metoden er av tilstrekkelig kvalitet med hensyn til senere bruk av materialet. Det ble sett på ivaretakelsen av integritet og anvendbarhet, mulighet for forvaltning i et langtidsperspektiv, hvor lett det er å gjenfinne informasjon i uttrekkene og hvorvidt det er mulig å tilgjengeliggjøre uttrekkene. Som en del av dette ble det spesifisert test caser fra seksjonene Bruk og Tilgjengeliggjøring basert på de mest brukte søk og oppslag mot datasett fra journaler og sakarkiver.

Det har blitt gjort test av programvare for dokumentkonvertering, mulighet for migrering og kvalitetssikring av beskrivelsene i malsettet. Tabellen under oppsummerer testene som er utført og hvilket behov de dekker.

Tabell 3: Tester for å kvalitetssikre metoden

Behov	**Beskrivelse**	**Hvordan teste?**
Anvendbarhet	Vurdering av malsett	Skjønnsmessig vurdering av om malsettet er forståelig og godt nok dokumentert.
Integritet	Kan uttrekket migreres?	Åpne SIARD-filen i en database-viewer og konvertere til CSV
Fremfinning	Filreferanser	Lage script med oppslag I XMLen som sjekker om filene eksisterer, samt om det finnes filer som det ikke refereres til
Tilgjengeliggjøring	Brukertester	Lage søkbar DIP-versjon for brukertjenester

Brukertester

I tillegg har prosjektgruppen ved hjelp fra seksjonene Bruk og Tilgjengeliggjøring kommet frem til en rekke brukertester basert på de mest brukte søk og oppslag mot datasett fra journaler og sakarkiver:

- Slå opp på gitt årstall og saksnummer i sakarkivet.
  Resultat: En sak, vises med tilhørende journalposter (og dokumenter) – hierarki

- Naviger eller søk fram aktuelt tema i arkivnøkkelen og hent fram sakene innenfor dette.
  Resultat: Flere saker, vises med tittel og år/nummer og mulighet til å gå videre ned på journalposter og dokumenter
- Søk på enkeltord eller sammensetninger av ord i tittel på sak (fritekstsøk).
  Resultat: Saker, vises med tittel og år/nummer og mulighet til å gå videre ned på journalposter og dokumenter
- Søk etter navngitt saksbehandler eller saksansvarlig (enkeltord eller fullt navn) i saker.
  Resultat: Saker, vises med tittel og år/nummer og mulighet til å gå videre ned på journalposter og dokumenter
- Søk på enkeltord eller sammensetninger av ord i tittel på journalpost (fritekstsøk).
  Resultat: Journalposter, vises med tittel, nummer og tilhørende sak
- Søk etter navngitt avsender (enkeltord eller fullt navn) i journalposter.
  Resultat: Journalposter, vises med tittel, nummer og tilhørende sak
- Søk etter navngitt mottaker (enkeltord eller fullt navn) i journalposter.
  Resultat: Journalposter, vises med tittel, nummer og tilhørende sak
- Søk på enkeltord eller sammensetninger av ord i tittel på dokument (fritekstsøk).
  Resultat: Dokumentposter, vises med tittel og nummer og tilhørende journalpost og sak
- Alle de ovennevnte søkene bør kunne kombineres med årstall og evt. dato.
- Og en filtreringsmulighet a la den i eInnsyn.no (OEP) vil være nyttig.

For å utføre disse testene ble det laget en prototype DIP med et Python-script som tok ut de relevante tabellene fra SIARD-filen og gjorde disse søkbare i en lokal SQLite-database. Erfaringene med denne testen var at SIARD-filen kunne gjøres søkbar raskt og på en bedre måte enn dagens fremfinningsmetoder. Siden testen var en POC ble det kun demonstert fritekstsøk og dato-søk. På grunn av tidsbegrensninger ble de andre testene bare mappet opp, men det ble ikke definert søk. Det er i midlertidig ingenting teknisk som står i veien for noen av testene som har blitt foreslått.

Samlet sett tok prototypetesten 2 dager.

Funnene fra disse testene tyder på at fremfinning i SIARD-uttrekk kan løses på en enkel måte av depotinstitusjonen på tross av manglende spesialprogramvare. Med bedre koordinering med malsettet og en standardisert informasjonsmodell vil det være mulig å lage standardiserte visningspakker basert på SIARD-filen.

Kvalitetssikring av uttrekket

For å se på kvaliteten av de faktiske uttrekkene vi har mottatt, tilsvarende mottakskontrollen som utføres ved avlevering til depot, har et utvalg tester blitt definert. Dette inkluderte test av dokumenter og filreferanser, samt kvalitetssikring av beskrivelsene i malsettet og metadataanalyse med hovedfokus på kompletthet. Tabellen under oppsummerer testene som er utført og hvilket behov de dekker.

Tabell 4: Tester for å kvalitetssikre uttrekket

Behov	**Beskrivelse**	**Hvordan teste?**
Integritet	Test av dokumenter	Filtypesjekk før og etter PDF/A-konvertering og PDF-validering etter konvertering
Fremfinning	Test av filreferanser	Lage script med oppslag I XMLen som sjekker om filene eksisterer, samt om det finnes filer som det ikke refereres til
Fremfinning	Kompletthetstester	Script som går igjennom XMLen og produserer opp analysetall.
Anvendbarhet	Vurdering av tilhørende dokumentasjon	Skjønnsmessig vurdering med tanke på fremtidig anvendbarhet av materialet
Anvendbarhet	Vurdering av kompabilitet mellom malsettet uttrekket?	Bruk decom og manuelt gå igjennom tabeller som har beskrivelse og sjekke at feltene er identiske i begge uttrekkene
Anvendbarhet og forvaltning	Migrering til annen database	Kontroll av at alle verdiene i primærnøkkelfeltene er unike og kontroll av at verdiene i fremmednøkkelfeltene samsvarer med verdier som forekommer i primærnøkkelfelter (eller andre kandidatnøkkelfelter) i de tabellene det refereres til
Test av format	Test av gyldig SIARD	På grunn av problemer med valideringsprogrammet KOSTVAL, må vi bruke migreringen til annen database som test for gyldig SIARD-format.

Kompletthetstester

For å sikre at datasettet vi har mottatt er komplett ble det utført ulike kompletthetstester. De ble definert med utgangspunkt i DIFI sin spesifikasjon som omhandler kvalitet på datasett i en datakatalog[8]. Dette har resultert i følgende tester:

- Opptelling av filreferanser sammenlignet med antall filer i uttrekket
- Kontroll av jevn fordeling av filer i arkivperiode (fremstilt som stolpediagram)
- Se hvilke filer i uttrekket som mangler filreferanser
- Sammenligne registerdata mot kildedata eller annet (komplett og korrekt) register

Konvertering til Noark 4

I henhold til riksarkivarens forskrift foreligger det krav om at uttrekk fra Noark 4 skal produseres i det avleveringsformat som er spesifisert i Noark versjon 4.1. For å imøtekomme dette blir Noark 4-uttrekk fremstilt basert på dataene i SIARD-versjonene.

Prosjektet jobber med å konvertere filene og de vil bli testet og godkjent i tråd med Arkivverkets rutiner for Noark 4-uttrekk.

Arbeidet så langt har vist at man i visse tilfeller trenger mer informasjon enn det man finner i basen for å fremstille Noark-versjonen. Dette kan eksempelvis være for informasjon som vises som 1 eller 0 i basen og forandres til andre verdier av applikasjonslogikken i selve Doculive-programvaren. Det viser at der vil være behov for en dialog med arkivskaper for å sikre at alt materialet kan tolkes.

Funn og vurdering

Her drøftes funnene fra piloten sett opp mot suksesskriteriene, samt at det gjøres en generell vurdering av metoden basert på funnene i piloten.

Ivaretakelsen av de arkivfaglige egenskapene autentisitet, integritet, anvendbarhet og pålitelighet

I avsnittet “Vurdering av metoden opp mot arkivfaglige egenskaper” i hoveddelen ble det sett på i hvilken grad metoden ivaretar autentisitet, integritet, anvendbarhet og pålitelighet. Det understrekes her at anvendbarhet er det mest relevante for denne metoden, da grunnlaget for de andre egenskapene legges tidligere i verdikjeden. I piloten har man derfor i tillegg testet hvor anvendbart materialet er.

[8] https://doc.difi.no/data/kvalitet-pa-datasett/#Komplett

Funnene fra piloten understøtter antagelsen fra hovedrapporten om anvendbarhet, materialet finnes og er teknisk tilgjengelig. Arbeidet har også vist hvor viktig det er med tilstrekkelig beskrivelser av databasestruktur, bruksmønster og andre sammenhenger som trengs for at materialet skal være forståelig og mulig å tolke.
Det anbefales å lage en veiledning med retningslinjer for hvilken type supplerende informasjon som må samles inn.

Forvaltning i et langtidsperspektiv

I piloten har vi testet forvaltning av de enkelte SIARD-uttrekkene ved enkel migrering fra SIARD-formatet til mySQL-format.

Et annet perspektiv er forvaltning av SIARD-formatet. At det er flere brukere av SIARD internasjonalt øker også muligheten for videreutvikling av både formatet og verktøy knyttet til dette. Dette gjør Arkiv-Norge bedre istand til å nyttegjøre oss av utvikling som foregår internasjonalt.

Fremfinning

Testene som omhandlet filreferanser viser at fremfinning av dokumenter og konkrete søk er mulig direkte i SIARD-uttrekket. Dette ble gjort ved å definere opp en enkel formidlingsversjon av uttrekket og legge opp til søk basert på de mest brukte søk i forbindelse med saksbehandling i brukerseksjonen.

Ny forståelse av administrativ bruk skaper behov for DIP

I forbindelse med ny forståelse av administrativ bruk er det forventet at Arkivverket vil overta råderetten for et stort antall av arkivene allerede ved mottak. Det er derfor essensielt at produksjonslinjen inkluderer hensiktsmessige DIP-versjoner som kan brukes til søk og gjenfinning av saksbehandlere i brukertjenester.

En måte å lage slike DIP-versjoner er i form av en sannsynlighetsbasert tilnærming. Ved å lage en enkel kopi av databasen med utgangspunkt i de mest brukte søkene vil saksbehandlere i brukertjenester i de aller fleste tilfeller kunne gjøre saksbehandlingen på egen hånd. Dersom det mot formodning skulle komme forespørsler som går utover dette må det samarbeides med personer med teknisk kompetanse som går inn i databasen og finner de relevante koblinger som trengs for å gjenfinne informasjonen. For at dette skal fungere må all informasjonen som er bevaringsverdig være i AIPen.

Tilgjengeliggjøring av uttrekkene

Det er ikke gjort noen konkrete tester med hensyn til tilgjengeliggjøring i piloten. Likevel understøtter funnene og en vurdering av formatet at tilgjengeliggjøring er mulig.

En av SIARD-databasene fra piloten ble brukt som case i en konseptstudie gjennomført i Arkivverket som ble kalt felles informasjonsmodell. Her utforsket man blant annet hvordan

gjenbruk av informasjonsmodeller kan redusere arbeidsmengde samt lette gjenfinning. Studien ble gjennomført ved å koble informasjonsmodeller av datasett til en felles informasjonsmodell Arkivverket har laget i forbindelse med studien. For datasettet fra SIARD-databasen ble de relevante koblingene som var nødvendige for mappingen identifisert, mens selve mappingen ble ikke utført. Dette på grunn av tidsbegrensning i studien. Det ble likevel konkludert med at mappingen ville vært mulig å gjennomføre på lik linje som for de andre datasettene. Studien viste at det er mulig å knytte datasett med sine respektive informasjonsmodeller til en felles informasjonsmodell og at sammenkoblingen mellom informasjonsmodell til et innlevert datasett og en felles informasjonsmodell var mindre ressurskrevende enn antatt.

Generelle vurderinger av produksjonslinjemetoden

Ved å ta eierskap til den delen av prosessen som normalt utføres av en uttrekksleverandør får depot mer kontroll over dette prosesstrinnet. Det medfører at vi kan være sikre på hvor autentisiteten blir ivaretatt og samtidig dokumentere hvor den ikke er mulig å ivareta. Det kan argumenteres mot metoden ved at depot påtar seg mer ansvar ved å overta denne jobben, og i verste fall blir det leddet som fører til brudd på autentisitet dersom man skal lage et Noark-uttrekk basert på SIARD-versjonen. Metoden fører samtidig til at depot får økt arbeidsmengde, særlig med tanke på arbeidet opp mot arkivskaper og innhenting av dokumentasjon rundt systemimplementasjon og bruksmønster. Dersom metoden skal fungere er det derfor viktig at gode retningslinjer for dette arbeidet utarbeides. Det kan også være en god idé å ta inn personell fra seksjonene Bruk og Tilgjengeliggjøring allerede på dette stadiet, da det er de som har best forutsetning for å vite hvilken informasjon som blir viktigst for gjenbruk.

Sannsynligheten for at feil oppdages og følges opp antas å være større enn om en uttrekksleverandør gjør jobben, ettersom en depotinstitusjon arbeider ut fra et annet formål enn en uttrekksleverandør. Depotinstitusjonen ønsker at informasjonen skal være tilgjengelig og anvendbar i et langtidsperspektiv, mens en uttrekksleverandørs hovedfokus er å lage et uttrekk som går igjennom testprogrammet og dermed godkjennes. Ved å ta eierskap til dette prosesstrinnet har depotinstitusjonen kontroll på dokumentering av konvertering, hvilke programmer som er brukt, samt logger og rapporter om evt feil. Man får også mulighet til å ta i bruk testverktøy og -metoder som vi ikke har hatt mulighet til å bruke før, eksempelvis identifisering av pronomtyper og test på filnivå før og etter konvertering.

Sett opp mot suksesskriteriene til piloten, samt føringene som ble lagt i MAVOD-rapporten vurderes metoden som et godt alternativ til dagens metoder.

Vedlegg 3: Tillegg til malsettet

Informasjon om Doculive-malen og Doculive-databaser

Doculive-databaser kan inneholde flere ulike datatyper og informasjon, i databasene til politiet er det Noark4 arkivdata, men databasene har også tabeller for plan/byggesak, Møte/utvalg, Saksgang og generelle dokumenter. Disse tabellene er ikke beskrevet i detalj, men i tabellbeskrivelsen er følgende betegnelser benyttet:

[DOCULIVE-BASIC] – generelle dokumenter (ikke-Noark4)

[BYGGESAK] – plan/byggesak

[SAKSGANG] – saksgang/workflow

[UTVALG] – Møte/utvalg (ikke funnet i de 3 pilotdatabasene)

Det er i tillegg en god del tabeller som ikke er relevante, disse er merket med [DUMMY], mange tabeller inneholder også system-informasjon og er merket med [SYSTEM]. Tabellene er ellers sortert etter prioritet med mere

Doculive-databasene er relasjonsdatabaser som er bygd opp uten skranker og fremmednøkler, med et sett av sentrale nøkler/indekser som definerer relasjonene mellom tabellene, men som ikke er åpenbare i strukturen.

Sentrale nøkkel-kolonner er saksnr, journalnr/jnr, objtype, objid1, objid2,objid3, elobjid/eldokid og edkid.

Av disse er saksnr og journalnr/jnr enklest å forklare, alle tabeller med kolonnen saksnr inneholder saksinformasjon og kan knyttes til hovedtabellen OA_SAK. Alle tabeller med kolonnen journalnr eller jnr inneholder journalpostinformasjon, og kan knyttes til hovedtabellen OA_DOKUMENT.

De andre nøkkel-kolonnene objtype,objid1,objid2,objid3, elobjid/eldokid og edkid, er brukt på ulike datatyper, og i Doculive-malen er disse merket med [NOTE-001] og [NOTE-002]

[NOTE-001] i Description i malen gjelder kolonner av type objtype,objid1,objid2,objid3, finnes i følgende tabeller:

- DL_COMMENT
- DL_LOG
- DL_CROSSREF
- EDOKTAB

OBJTYPE bestemmer betydningen av OBJID1,OBJID2 og OBJID3. Jeg har beskrevet de verdiene av OBJTYPE som finnes i politi-databasene, Doculive-databaser med f eks Møte/utvalgsdata har andre verdier enn disse.

DL_COMMENT | table23 {n4:MERKNAD} {n5:merknad} Merknad
- Objtype 449 => Saksmerknad, OBJID1 = OA_SAK.SAKSNR
- Objtype 450 => Journalpostmerknad, OBJID1 = OA_DOKUMENT.JOURNALNR

DL_LOG | table48 {n4:Tilleggsinfo} {n5:endringslogg} Tilleggsinformasjon, aktivitetslogg
- Objtype 449 => Saksaktivitet, OBJID1 = OA_SAK.SAKSNR

- Objtype 450 => Journalpostaktivitet, OBJID1 = OA_DOKUMENT.JOURNALNR
- Objtype 470 => Dokument/filaktivitet, OBJID1 = EDOKTAB.OBJID1 = CA_ELDOK.ELDOKID=DL_ELDOKLINK.ELOBJID

DL_CROSSREF | table26"{n4:JFSAK} Jevnfør sak, også referanser til og mellom journalposter",

- Fromobjtype 449 => Referanse fra sak, FROMOBJID1 = OA_SAK.SAKSNR
- Fromobjtype 450 => Referanse fra journalpost, FROMOBJID1 = OA_DOKUMENT.JOURNALNR
- Toobjtype 449 => Referanse til sak, TOOBJID1 = OA_SAK.SAKSNR
- Toobjtype 450 => Referanse fra journalpost, TOOBJID1 = OA_DOKUMENT.JOURNALNR

EDOKTAB | table102 [DIP] {n4:DOKVERS} {n5:dokumentobjekt} Dokumentversjon [NOTE-002] .... Fil-link og informasjon, se også DL_ELAREA og CA_ELDOK

- Objtype 480 => Fil tilknyttet noark4-arkivet, OBJID1 = DL_ELDOKLINK.ELOBJID=CA_ELDOK.ELDOKID, se også [NOTE-002]

Fil-lager, fil-informasjon og fil-link/peker i Doculive databaser

[NOTE-002] i Descripton i malen er satt inn på alle tabellene med som inneholder slik informasjon.

DL_ELAREA | table32 {n4:LAGRENHET} Lagringsenhet elektronisk arkiv

- Beskriver fil-arkivene i databasen, det kan være fra ett og oppover. Det er 3 typer, definert i kolonnen MEDIUM. Doculive kan være satt opp med flere fil-arkiv, også av ulike typer.
- 0 = fil-arkiv på lokal mappe eller share, brukt i politi-databasene til politiet
- 2 = fil-arkiv på ftp-server, Pilot-databasene til politiet benytter dette som hovedløsning
- 3 = fil-arkiv i databasen, filene ligger i tabellen edokfiles, kolonne = efile (oftest type BLOB eller LONG RAW)

Eksempler på innhold i DL_ELAREA for de tre hovedtypene:

Medium= 0, Fil-arkiv på lokal mappe eller share. Serverpath angir vanligvis startmappe/share, blank her, da ligger mappeinformasjonen i tabellen edoktab, se beskrivelse under

pathid (INT)	name (VARCHAR)	medium (INT)	description (VARCHAR)	dbtable (VARCHAR)	server (VARCHAR)	serverpath (VARCHAR)	username (VARCHAR)	password (VARCHAR)	dirdelim (CHAR)	sndsize (INT)	rcvsize (INT)	numdoc (INT)	n1 (INT)	n2 (INT)	n3 (INT)	n4 (INT)	n5 (INT)	n6 (INT)	cachepid (INT)	updated (TIMESTAM
10	dokumenter	0			\u005c\u005carkiv15\u005cdl_data							86392	111	118	120	101	97	97		2002-10-15T07:03:12.770000Z

Medium = 2, Fil-arkiv på ftp server. Server, bruker og passord ligger direkte i tabellen (!) Startmappe for filarkivet på ftp-serveren ikke mulig å se her, mappen blir default startmappe for brukeren innlogget på ftp-serveren. Ftp-site = serverpath

pathid (INT)	name (VARCHAR)	medium (INT)	description (VARCHAR)	dbtable (VARCHAR)	server (VARCHAR)	serverpath (VARCHAR)	username (VARCHAR)	password (VARCHAR)	dirdelim (CHAR)	sndsize (INT)	rcvsize (INT)	numdoc (INT)	n1 (INT)	n2 (INT)	n3 (INT)	n4 (INT)	n5 (INT)	n6 (INT)	cachepid (INT)	updated (TIMESTAM
1	FtpHist15	2	Fillager\u005c\u005cNasC\u005cDoculive\u005c15-Hist		ARKIV-PROD-004	Histarkiv15.ftpHist15	POLITIET\u005cftpdl	Ftpuser01	/			184082	112	112	109	107	97	97		2018-06-19T22:00:00.000000Z

Filarkiv i databasen:

PATHID (INT)	NAME (VARCHAR)	MEDIUM (INT)	DESCRIPTION (VARCHAR)	DBTABLE (VARCHAR)
2	database	3		edokfiles

EDOKTAB | table102 [DIP] {n4:DOKVERS} {n5:dokumentobjekt} Dokumentversjon [NOTE-002] ....
Fil-link og informasjon, se også DL_ELAREA og CA_ELDOK

Eksempel på viktige kolonner i edoktab fra politi-pilot:

pcpath {VARCHAR}	pcfil {VARCHAR}	pcfiltype {VARCHAR}	pathid {INT}	subpath {VARCHAR}	fil {VARCHAR}	filnr {VARCHAR}	filtype {VARCHAR}
D:\u005cdl_fileload\u005cupload\u005c	PMS002-101220021	doc	1	/dir00000	00000002		doc

Pcpath = lokal mappe før arkivering i Doculive

Pcfil = filnavn før arkivering

Pathid = dl_elarea.pathid, fil-arkiv ident/indeks

Subpath = mappe/mapper under start-mappe gitt i dl_elarea, både \ og / brukes som mappeskilletegn.

Fil + filnr + . + filtype = filnavn i Doculive fil-arkivet

I eksempelet over finnes filen på \dir00000\00000002.doc

Det er svært ulike sammensetninger av mappestrukturen i dl_elarea og edoktab

Vedlegg 4: Testrapport Nordre Buskerud politidistrikt

Dette avsnittet oppsummerer resultatene fra testene som ble kjørt på SIARD-filen som inneholder databasen fra Doculivesystemet til Nordre Buskerud politidistrikt.

Test av gyldig SIARD

Per i dag eksisterer det få verktøy som validerer mot SIARD-standarden. Det ble forsøkt å validere uttrekket (både før og etter bearbeidelse) via KOST-val[9], men det lot seg ikke validere av dette verktøyet. På grunn av manglende beskrivelser i feilmeldingen ble ikke dette fulgt opp videre i piloten.

Ettersom SIARD-uttrekket lot seg behandle og migrere tilbake til mySQL-database ble det besluttet at uttrekket kan godkjennes uten denne valideringen. Det bør testes på nytt når bedre verktøy er tilgjengelige.

Test av dokumenter

Filtypesjekk før og etter PDF/A-konvertering

Tabellene under viser antall filer, pronomkode og filtype før og etter konvertering.

Etter konvertering		
Antall filer	**Pronomkode**	**Filtype**
24895	fmt/95	Acrobat PDF/A - Portable Document Format
17653	fmt/354	Acrobat PDF/A - Portable Document Format
17545	fmt/19	Acrobat PDF 1.5 - Portable Document Format
16617	x-fmt/111	Plain Text File
14955	fmt/16	Acrobat PDF 1.2 - Portable Document Format
13516	fmt/18	Acrobat PDF 1.4 - Portable Document Format
4247	UNKNOWN	
1111	fmt/17	Acrobat PDF 1.3 - Portable Document Format
794	fmt/43	JPEG File Interchange Format
671	fmt/12	Portable Network Graphics
285	fmt/20	Acrobat PDF 1.6 - Portable Document Format
100	fmt/11	Portable Network Graphics
85	fmt/276	Acrobat PDF 1.7 - Portable Document Format
66	fmt/44	JPEG File Interchange Format
40	fmt/13	Portable Network Graphics
38	fmt/42	JPEG File Interchange Format
36	fmt/41	Raw JPEG Stream
25	fmt/134	MPEG 1/2 Audio Layer 3
23	fmt/353	Tagged Image File Format
8	fmt/157	Acrobat PDF/X - Portable Document Format - Exchange 1a 2001
8	fmt/15	Acrobat PDF 1.1 - Portable Document Format
6	fmt/488	Acrobat PDF/X - Portable Document Format - Exchange PDF/X-4
1	fmt/199	MPEG-4 Media File
1	fmt/14	Acrobat PDF 1.0 - Portable Document Format
1	fmt/158	Acrobat PDF/X - Portable Document Format - Exchange 3 2002

[9] http://www.preforma-project.eu/kost-val.html

Før konvertering		
Antall filer	**Pronomkode**	**Filtype**
17847	fmt/40	Microsoft Word Document
17653	fmt/354	Acrobat PDF/A - Portable Document Format
17545	fmt/19	Acrobat PDF 1.5 - Portable Document Format
15868	x-fmt/430	Microsoft Outlook Email Message
14955	fmt/16	Acrobat PDF 1.2 - Portable Document Format
13516	fmt/18	Acrobat PDF 1.4 - Portable Document Format
5043	x-fmt/111	Plain Text File
4553	fmt/412	Microsoft Word for Windows
1111	fmt/17	Acrobat PDF 1.3 - Portable Document Format
794	fmt/43	JPEG File Interchange Format
699	UNKNOWN	
680	fmt/61	Microsoft Excel 97 Workbook (xls)
671	fmt/12	Portable Network Graphics
642	fmt/214	Microsoft Excel for Windows
464	fmt/95	Acrobat PDF/A - Portable Document Format
286	fmt/598	Microsoft Excel Template
285	fmt/20	Acrobat PDF 1.6 - Portable Document Format
202	fmt/4	Graphics Interchange Format
118	fmt/126	Microsoft Powerpoint Presentation
102	fmt/50	Rich Text Format
100	fmt/11	Portable Network Graphics
96	fmt/96	Hypertext Markup Language
85	fmt/276	Acrobat PDF 1.7 - Portable Document Format
85	fmt/645	Exchangeable Image File Format (Compressed)
74	fmt/215	Microsoft Powerpoint for Windows
66	fmt/44	JPEG File Interchange Format
64	x-fmt/391	Exchangeable Image File Format (Compressed)
40	fmt/13	Portable Network Graphics
38	fmt/42	JPEG File Interchange Format
36	fmt/41	Raw JPEG Stream
25	fmt/134	MPEG 1/2 Audio Layer 3
23	fmt/353	Tagged Image File Format
15	fmt/99	Hypertext Markup Language
15	x-fmt/266	GZIP Format
14	x-fmt/263	ZIP Format
9	fmt/116	Windows Bitmap
9	fmt/563	Adobe Illustrator
8	fmt/157	Acrobat PDF/X - Portable Document Format - Exchange 1a 2001
8	fmt/15	Acrobat PDF 1.1 - Portable Document Format
7	fmt/395	vCard
6	fmt/488	Acrobat PDF/X - Portable Document Format - Exchange PDF/X-4
6	fmt/3	Graphics Interchange Format
6	fmt/597	Microsoft Word Template
6	x-fmt/45	Microsoft Word Document Template
3	fmt/445	Microsoft Excel Macro-Enabled
3	x-fmt/390	Exchangeable Image File Format (Compressed)
3	fmt/291	OpenDocument Text
2	fmt/657	Open XML Paper Specification
2	fmt/53	Rich Text Format
1	fmt/111	OLE2 Compound Document Format
1	fmt/487	Macro Enabled Microsoft Powerpoint
1	fmt/199	MPEG-4 Media File
1	fmt/523	Macro enabled Microsoft Word Document OOXML
1	fmt/290	OpenDocument Text
1	fmt/754	Microsoft Word Document (Password Protected)
1	fmt/14	Acrobat PDF 1.0 - Portable Document Format
1	fmt/583	Vector Markup Language
1	fmt/158	Acrobat PDF/X - Portable Document Format - Exchange 3 2002
1	fmt/355	Rich Text Format
1	x-fmt/384	Quicktime

I testen gikk antall filer identifisert som UNKNOWN opp, dette er fordi en del tekstfiler produsert av konverteringen av Outlook-meldinger fikk denne statusen av testverktøyet. Det er gjort stikkprøver på filene og disse kan leses som vanlige tekstfiler.

Det totale antall filer før og etter konvertering er ulikt. Dette kommer av at enkelte filer ikke lot seg konvertere, enten fordi de var i formater som ikke kunne konvertere til PDF/A eller var passordbeskyttet. Disse filene er gjennomgått og de er ikke bevaringsverdige filer. Imidlertid har metoden tatt vare på samtlige filer fra før konverteringen, i tillegg til de konverterte filene.

Kontroll av PDF/A

Det ble utført PDF/A-kontroll på filene ved bruk av VERA PDF[10]. Kontrollen viser at alle filene som ble konvertert som en del av piloten var gyldige PDF/A.

Kompletthetstester

Opptelling av filreferanser sammenlignet med antall filer i uttrekket

Filreferansene ble sjekket mellom SIARD-tabellen med referanser til dokumentene og dokumentmappen. Sjekken ble gjort fra begge sider først for å finne ut om det var filreferanser i SIARD-tabellen der filene manglet og hvorvidt det lå filer i dokumentmappen som manglet filreferanse fra SIARD-tabellen. Sammenligningen viste at:

- 113899 filer matchet referansen fra SIARD-filen
- 1 fil matchet ikke referansen fra SIARD-filen
- 35 filer hadde ikke oppføring i SIARD-filen

Den manglende filreferansen tilhører mest sannsynlig et dokument med nesten samme navn i dokumentmappen. Filene som manglet filreferanse er trolig slettede filer eller kladder som har blitt liggende igjen i systemet.

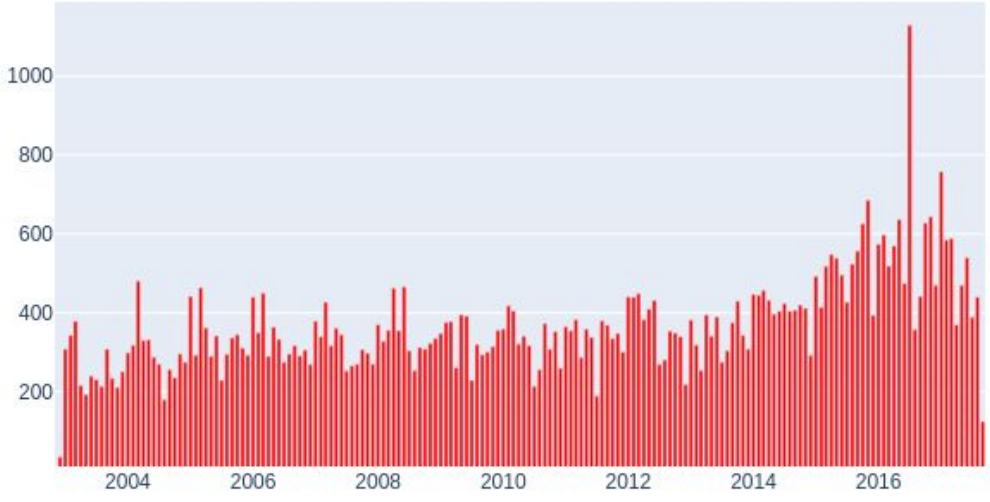
[10] https://verapdf.org/

Fordeling av filer i arkivperiode

Tabellen viser fordeling av saksposter og journalposter per år.

År	Antall saksposter	Antall journalposter
1890		1
1941		1
1988		1
1992		2
1993		1
1994		6
1995		2
1996		6
1997		2
1998		10
1999		7
2000		9
2001		15
2002	1	56
2003	1500	3128
2004	1797	3561
2005	1782	4003
2006	1815	3981
2007	1714	3833
2008	1722	4174
2009	1604	3965
2010	1721	3925
2011	1649	4009
2012	1732	4367
2013	1709	4116
2014	1921	4941
2015	2889	6222
2016	3315	7044
2017	1962	4265
mangler år	2	

Figuren under inneholder fordeling av journalposter per måned. Den viser en jevn og troverdig fordeling, som tyder på at datasettet er uten store mangler.



Vurdering av tilhørende dokumentasjon

I tillegg til beriket SIARD-fil har politiet gitt dokumentasjon som kan være nyttig med tanke på gjenbruk av materialet. Å ikke ha med tilstrekkelig tilhørende dokumentasjon er identifisert som noe av det mest kritiske med metoden.

Dokumentasjonen som skal supplere omfatter:

- Bruksmåter (som beskriver omfanget og hvilken informasjon som skal i DocuLive)
- Beskrivelse av utvalgte rutiner og enkelte sakstyper
- Arkivnøkkel, der innledende beskrivelser vil være av interesse
- Standard for arkiv og arkivdeler i Doculive
- Standard for registrering av anskaffelser

Vurdering av kompatibilitet mellom malsett og uttrekket

En manuell gjennomgang av malsettet viser at tabellene sammenfaller med malsettet og det ble konkludert med at det ikke var behov for ytterlige suppleringer.

Alle feltene i tabellene som var markert som high og medium priority har beskrivelser og det ble vurdert at de var mulige å tolke.

Test av migrering

Migrering ble testet ved å konvertere SIARD-versjonen til en mySQL-database med verktøyet SCFC.

Det ble samtidig utført en kontroll av at alle verdiene i primærnøkkelfeltene er unike, samt at fremmednøkkelfeltene samsvarer med verdier som forekommer i primærnøkkelfelter i de tabellene det refereres til.