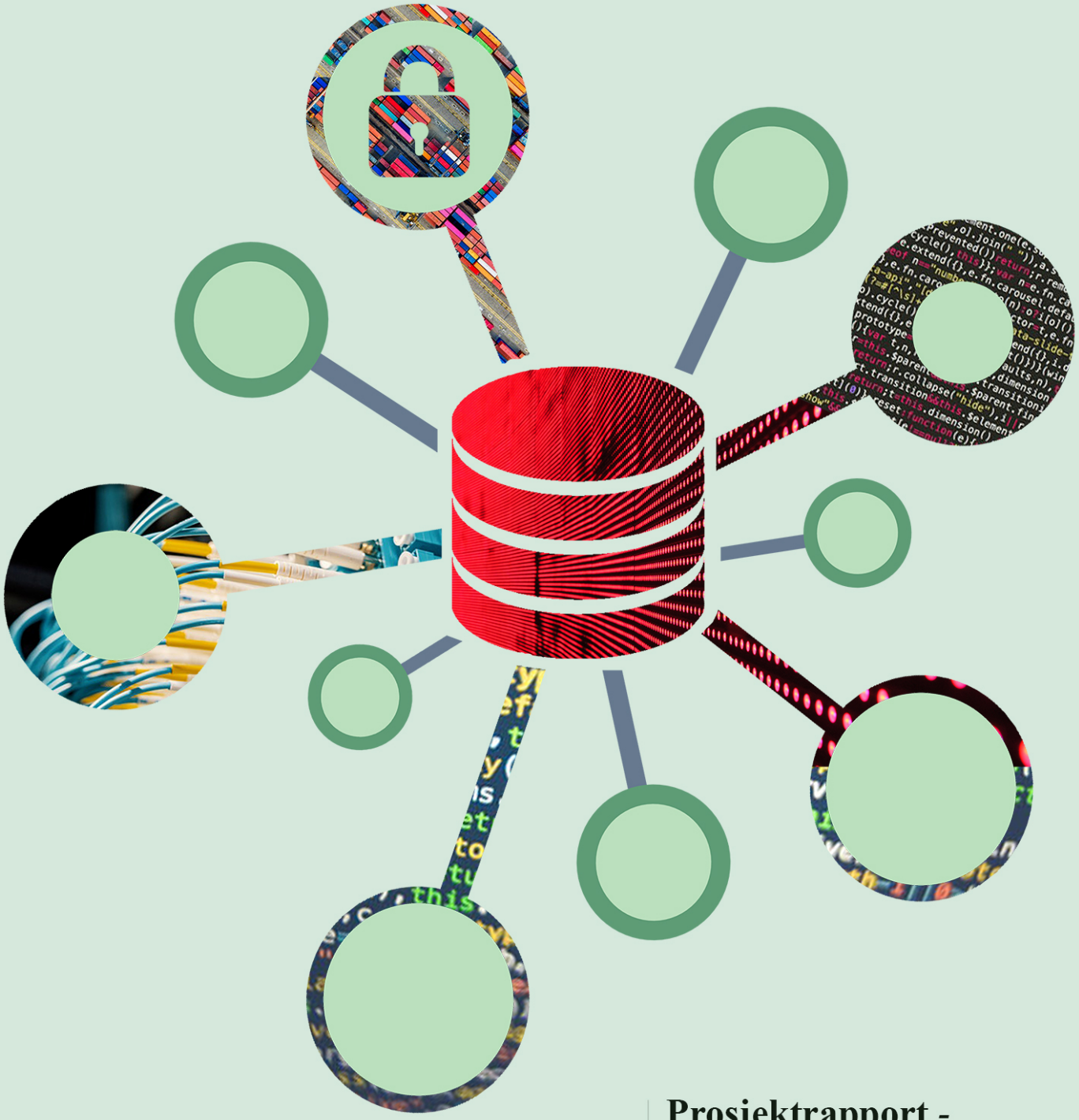


Prosjektrapport - Metode for bevaring og tilgjengeliggjøring av digitalt skapt privatarkiv

Stig Brunstad
Morten Eek
Frode Kirkholt

27.01.2021

Innhold

Kapittel 1

Innledning

Dette prosjektet kom i stand for å utvikle og beskrive en helhetlig metode for å bevare og tilgjengeliggjøre digitalt skapt privatarkiv. Målet var å legge til rette for at flere digitalt skapte privatarkiv kunne bli bevart og gjort tilgjengelig. I denne rapporten beskriver vi resultatene av prosjektet, og inkluderer også en oppsummering av metode for utvikling av bevaringsplan. Denne er utviklet av andre.

Prosjektets arbeid og hovedvekten av det som presenteres i denne rapporten er knyttet til den mer tekniske siden av bevaring og tilgjengeliggjøring av digitalt skapt privatarkiv. Vi håper at også andre enn de som allerede arbeider med digitalt skapt arkiv leser denne rapporten. En del av formålet med rapporten er å gjøre bevaring og tilgjengeliggjøring av digitalt skapt privatarkiv enkelt å forstå, ikke bare metodisk, men også hva som kreves for å kunne komme i gang med bevaring *og* tilgjengeliggjøring av vår felles kulturarv.

1.1 Er det behov for en ny metode?

I prosjektet har vi ved flere anledninger stilt oss selv spørsmålet om det er behov for en egen metode for bevaring av digitalt skapt privatarkiv. Hver gang blir svaret det samme. Hvis vi ser bort fra det å utarbeide bevaringsplan for privatarkiv er det, fra et teknisk ståsted, lite som skiller bevaring og tilgjengeliggjøring av digitalt skapt privatarkiv og digitalt skapte arkiv fra offentlig sektor. Dette har vi forholdt oss til i prosjektet. Det har resultert i en metode som kan passe alt digitalt skapt arkiv. Privatarkiv har noen trekk som til dels skiller det fra arkiv fra offentlig sektor. Disse ivaretas og beskrives, men inngår stort sett sømløst i den helhetlige tilnærmingen til bevaring og tilgjengeliggjøring i metoden som beskrives i rapporten.

På den annen side er det åpenbart at noe må gjøres for å øke antallet bevarte og tilgjengeliggjorte privatarkiv. I tall innsamlet av Arkivverket for 2019[1] fra arkivinstitusjoner, bibliotek, lokalhistoriske arkiv og museer som har tatt i mot digitalt skapt privatarkiv, er det totale antallet mottatte avleveringer og deponeringer på 454. Litt forsiktig antydes det i rapporten at dette er et relativt lite antall i forhold til alt som er skapt. Videre antydes det at deler av de mottatte digitalt skapte privatarkivene ikke er tilstrekkelig sikret. Selv om det ikke fremkommer fra tallene som presenteres, er det rimelig å anta at det mottatte digitalt skapte materialet heller ikke fremstår som anvendelig eller er gjort tilgjengelig for hverken arkivarer eller publikum. Det er verdt å minne om at dette også er tilfelle i kommunal sektor. Der er det store etterslep på avlevering av digitalt skapt arkiv, og svært lite av det som er avlevert er tilgjengelig for brukere.

Det er iverksatt enkelte tiltak for å bidra til at flere privatarkiv, og da også digitalt skapte privatarkiv, blir bevart og gjort tilgjengelig. Noen av disse har gitt gode resultater for privatarkivarbeidet. Gjennom utarbeidelse av bevaringsplaner pågår det blant annet en omfattende kartlegging av hvilke privatarkiv som er samlet inn, og hvilke som bør samles inn for å bidra til å sikre kulturarven. Dette vil indirekte også omfatte digitalt skapt privatarkiv. Det er også en satsing på privatarkiv i visjonen for det nye Digitalarkivet.[2] Dette

[1] Tall og statistikk: https://www.arkivverket.no/arkivutvikling/tall-og-data-statistikk/arkivstatistikken-amb/tal-og-analyser-fra-2019

[2] Det nye Digitalarkivet: https://www.digitalarkivet.no/content/nytt-digitalarkiv

er et prosjekt som foregår i regi av Arkivverket. I innledningen til Arkivverkets bevaringsplan for privatarkiv[3] skriver Riksarkivaren at Arkivverkets mål er at flest mulig privatarkiv skal være tilgjengelig i Digitalarkivet. I tillegg er det planer om at Digitalarkivet skal tilby ulike tjenester for lagring og tilgjengeliggjøring av digitalt skapt arkiv, men det er et stykke frem før dette vil være tilgjengelig.

I mellom disse to tiltakene, metode for bevaringsplan og fremtidige nye tjenester i Digitalarkivet, befinner det arbeidet seg som må utføres på kildene som skal bevares og tilgjengeliggjøres. Det er oppgaven som tilfaller alle oss som tar i mot eller skal ta i mot, tilgjengeliggjøre og forvalte digitalt skapt privatarkiv. Metoden som beskrives i denne rapporten omfatter nettopp dette.

1.2 Om metoden

Metoden som presenteres i denne rapporten skiller seg på noen sentrale punkter fra andre metoder vi er kjent med, og som benyttes for bevaring og tilgjengeliggjøring av digitalt skapt arkiv i Norge i dag. Den er resultatene av en større endring i måten vi tenker på og behandler digitalt skapt arkiv ved Bergen byarkiv. Denne endringen har pågått de siste tre årene, og inkluderer også det prosjektet som presenteres her. I forkant av dette hadde vi gradvis blitt klar over at vi hadde store utfordringer knyttet til nesten alle sider ved bevaring av digitalt skapt arkiv. Metodene, standardene og verktøyene vi benyttet, kunne hverken gi oss en effektiv og god arbeidsprosess eller produsere et resultat som kunne sikre oss en tilfredsstillende langtids-tilgang til det materialet vi skulle forvalte.

Vi så at enkelte deler av arbeidet med bevaring, forvaltning og tilgjengeliggjøring av digitalt skapt arkiv isolert sett fungerte, mens andre var unødvendig tidkrevende og kompliserte. Ofte ville resultatene fra én del av prosessen fungere dårlig i neste del av prosessen som data skulle gjennomgå. Uavhengig av hvilken metode eller hvilke verktøy vi benyttet, endte vi opp med å måtte forvalte et materiale som ikke var tilstrekkelig kvalitetssikret. Dette fremsto som vanskelig å forvalte over tid og var relativt utilgjengelig. Standardene vi benyttet var i liten grad anerkjent eller utbredt utover et lite felt innen arkivfaget. Det var en svært liten del av materialet i depot som kunne gjøres tilgjengelig for bruk uten omfattende bearbeiding. Den lille delen av materialet i depot vi kunne gjøre tilgjengelig, var avhengige av bearbeiding i spesialisert verktøy, før det kunne forberedes for tilgjengeliggjøring. Det materialet vi hadde gjort tilgjengelig i vår innsynsløsning i Bergen kommune, var ikke basert på den versjonen av data som lå i depot, men hentet fra opprinnelig plassering og bearbeidet parallelt ved mottak av data, da dette var den eneste måten å sikre at vi kunne gjøre materialet tilgjengelig.

Det var åpenbart at vi måtte iverksette noen tiltak for å endre på dette komplekset av utfordringer. Det var også åpenbart at det ikke ville være tilstrekkelig å skulle fokusere på enkeltområder. Det var behov for en helhetlig tilnæring til bevaring, forvaltning og tilgjengeliggjøring av digitalt skapt arkiv.

Alle valg gjort under utvikling av metoden er gjort for å sikre at alt digitalt skapt arkiv som bevares, både kan forvaltes over tid og skal kunne gjøres tilgjengelig for brukere, nå og i fremtiden. Arkivene, eller dataene, skal med andre ord i størst mulig grad være uavhengig av verktøy og standarder som setter begrensninger på formålet med bevaring: det å sikre langtids tilgang til kildene. Metoden sikrer at arkivene som bevares er anvendbare i en helt annen skala enn ved bruk av andre metoder vi er kjent med. Den oppfyller også minst samme krav til sporbarhet i operasjoner som kan påvirke data når de forberedes for lagring i depot og skal forvaltes der.

1.3 Om verktøy

I denne rapporten har vi valgt å ikke bare beskrive en metode, men også inkludere beskrivelser av to verktøy som benyttes for å anvende metoden. Det er vårt syn at det ikke lenger er rasjonelt å skille metode og verktøy som benyttes ved arbeidet med digitalt skapt arkiv. Metode og verktøy er utviklet parallelt. Det har

[3] Bevaringsplan for privatarkiv i Arkivverket: https://www.arkivverket.no/for-arkiveiere/arkiver-fra-privat-sektor/arkivverkets-bevaringsplan-for-privatarkiv

foregått i en iterativ prosess basert på praktiske erfaringer i forbindelse med vårt arbeid med bevaring og tilgjengeliggjøring av digitalt skapt arkiv.

Vi beskriver også et tredje verktøy som ikke er utviklet hos oss. Til sammen dekker disse tre verktøyene de mest sentrale deler av det arbeidet som foregår direkte med data. Fra oppgaver ved tilrettelegging av data for langtids lagring i depot, dokumentasjon og forberedelse for tilgjengeliggjøring i innsynsløsning, og til sist overvåkning og forvaltning av data i digitalt depot.

Teknologi er uunnværlig for alt digitalt skapt arkiv, fra det dannes til det brukes av fremtidige generasjoner. Den samme uunnværlige teknologien og bruken av den vil være i kontinuerlig endring. Konsekvensen av endringer er behov for å forvalte og tilpasse metoden og verktøyene for å møte nye utfordringer. Det innebærer at de vil og må endres og utvikles over tid, men grunntrekkene og grunnfunksjonaliteten vil i stor grad forbli uforandret.

Selv om verktøy som benyttes ved bevaring og tilgjengeliggjøring har høy grad av automatisering fjerner ikke dette behovet for svært god IT-kompetanse. I arbeid med digitalt skapt arkiv vil det oppstå situasjoner og hendelser som krever oppfølging, og at man tar valg som påvirker data. Da er det viktig at man er i stand til å ta de rette valgene for å sikre langtids-tilgjengelighet til kildene. Like fullt er arbeid med digitalt skapt arkiv også avhengig av arkivkompetanse. Digitalt skapt arkiv beskrives på samme måte som annet arkiv i arkivkatalogen, og tilgjengeliggjøres for brukere. Både arkiv- og IT-kompetansen må trekke i samme retning, men innenfor sine respektive fagområder.

Arkivmiljøet i Norge har brukt lang tid på å skaffe seg IT-kompetanse. I kommunal sektor har det hos byarkiv, fylkesarkiv og interkommunale arkivtjenesteleverandører, vært en gradvis økning av personer med god IT-kompetanse. Noen av disse tar også imot digitalt skapt privatarkiv. Flertallet av disse er relativt store aktører hvor det har vært avgjørende å ha IT-kompetanse for å kunne følge opp lovkrav som styrer bevaringsarbeidet med offentlige arkiv. Privatarkivfeltet er i langt større grad preget av mange små aktører som bevarer privatarkiv. Det er heller ikke alle som nødvendigvis har bevaring av arkiv som sin primæroppgave. For noen av disse vil det være for ressurskrevende å skaffe seg tilstrekkelig IT-kompetanse. Dette er en situasjon som hverken metoder, verktøy eller nye tjenester i Digitalarkivet, kan gjøre noe med. For disse aktørene vil samarbeid og kjøp av tjenester være et svært godt alternativ for å sikre at mottatte og planlagt mottatte digitalt skapte privatarkiv blir bevart og kan gjøres tilgjengelig.

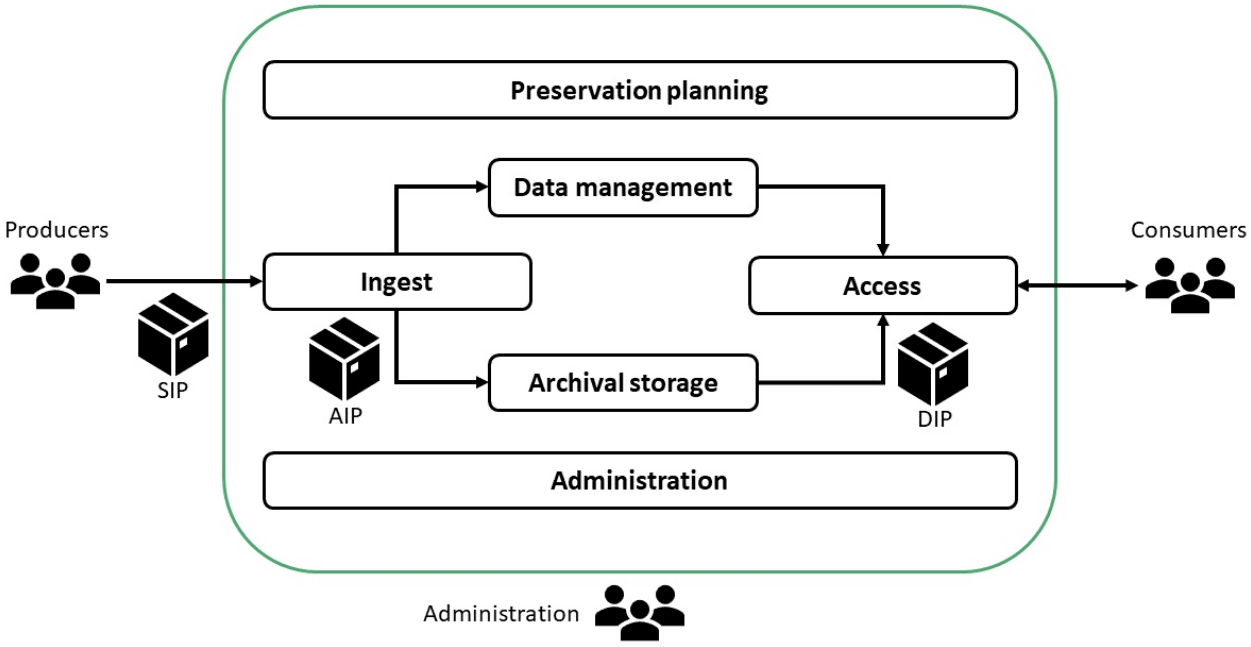
1.4 Teoretisk rammeverk

Den metoden og anvendelsen av den som presenteres i den resterende delen av rapporten, er bygget opp i henhold til referansemodellen OAIS – Open Archival Information System.[4] I denne sammenhengen referere begrepet «Information System» til de funksjonene som inngår i den tekniske, forvaltningsmessige og praktiske delen av bevaring og tilgjengeliggjøring av digitalt skapt arkiv.

I den forenklede illustrasjonen av OAIS over vises roller og funksjoner som inngår i referansemodellen. Hovedvekten av det som beskrives videre i denne rapporten er knyttet til arbeid med de tre pakkene SIP, AIP og DIP. SIP (submission information package) inneholder data som avleveres og hentes ut fra opprinnelig plassering. AIP (archival information package) inneholder data som er prosessert og dokumentert for å kunne lagres og forvaltes i digitalt depot. Og DIP (dissemination information package) inneholder data og dokumentasjon som er tilrettelagt for tilgjengeliggjøring.

Rapporten vil direkte og indirekte omtale funksjonene og pakkene i OAIS-referansemodellen som er vist i illustrasjonen.

[4] OAIS - Open Archival Information System: https://public.ccsds.org/pubs/650x0m2.pdf



Figur 1.1: OAIS funksjonsmodell

1.5 Om oppbygging av rapporten

Kapitlet "Planlegging av bevaring og innsamling" oppsummerer metode for utvikling av bevaringsplan for privatarkiv. I siste halvdel beskrives noen aspekt ved digitalt skapt privatarkiv som kan påvirke prioriteringer ved utforming og anvendelse av en innsamlingsplan.

Kapitlene "Uttrekk", "Normalisering" og "Dokumentasjon" er knyttet til *ingest*-funksjonen i OAIS-referansemodellen. Slik metoden som presenteres er lagt opp, inkluderer det en rekke steg i prosessen de digitalt skapte arkivene skal gå gjennom. SIP, AIP og DIP utformes som del av *ingest*.

Kapitlet "Tilgjengeliggjøring av digitalt skap privatarkiv" er knyttet til *access*-funksjonen i OAIS, og beskriver hvordan DIP og arkivkatalog tas i bruk for å gjøre arkiv tilgjengelig for brukergrupper.

I kapitlet "Forvaltning i depot" er det *archival storage*-funksjonen som beskrives, hvor AIP lagres, overvåkes og forvaltes.

Mot slutten av rapporten beskrives verktøy som er utviklet eller beskrevet for å støtte de funksjonene nevnt over. Kapitlene "Uttrekk", "Normalisering", "Dokumentasjon" og "Tilgjengeliggjøring" bør leses i sammenheng med kapitlet "Verktøy".

Både referansemodellen OAIS og metode og verktøy som beskrives her, danner rammene for hvordan bevaring og tilgjengeliggjøring av digitalt skapt privatarkiv kan og bør gjøres, men det er opp til hver enkelt bevaringsinstitusjon å sette dette inn i sin kontekst.

1.6 Om avvik fra søknad

I opprinnelig søknad ville vi benytte et arkiv for å vise metoden i bruk. Vi måtte gå bort fra den tilnærmingen da vi så at den ikke ville bidra til å styrke formålet med rapporten. Videre har vi gått bort fra å beskrive håndtering av spesifikke filformater. Vi anser dette for å inngå i den helhetlige metoden som beskrives, og være del av oppgavene som håndteres av verktøy som er beskrevet, eller inngå som forvaltning av data i depot over tid.

1.7 Prosjektorganisering

Styringsgruppen har bestått av:

- Karin Gjelsten, avdelingsleder ved Bergen byarkiv
- Terje Haram, daglig leder i ArkiVest - samarbeidspartner i prosjektet
- Randi Christine Sande, leder for programmet Digitalt fornyelse ved Bergen byarkiv

Prosjektgruppen har bestått av:

- Frode Kirkholt, Bergen byarkiv
- Morten Eek, Bergen byarkiv
- Stig Narve Brunstad, Bergen byarkiv

Prosjektet er gjennomført av Bergen byarkiv med ArkiVest som samarbeidspartner. Prosjektet er gjennomført med støtte fra Riksarkivaren.

Prosjektgruppen ønsker å takke Arkivverket for den hjelp vi har fått og den tålmodigheten de har hatt med et forsinket prosjekt.

Vi håper rapporten vil bidra til at metoden og verktøyene blir brukt slik at vi sammen kan bidra til å bevare mer av vår digitalt skapte kulturarv.

Kapittel 2

Planlegging av bevaring og innsamling

I privat sektor, og for stort sett alle mennesker i Norge, har overgangen til en digital hverdag gått fort. Overgangen har påvirket mange av våre daglige gjøremål, både på jobb, skole og i fritiden. Til tross for dette er mengden digital samfunnsdokumentasjon i den samlede arkivbestanden i Norge betydelig underrepresentert. De færreste bevaringsinstitusjoner som forvalter privatarkiv, har en tilstrekkelig representasjon i sin arkivbestand. Det er nok flere grunner til dette, men den underliggende årsaken er det digitale formatet i seg selv, og institusjonenes kunnskap om og erfaring med bevaring av digitalt skapt arkiv.

I St.meld. 7 (2012-2013)[1] ble det påpekt at bevarte privatarkiv var underrepresentert i forhold til offentlige arkiver, slik at det ikke fantes en helhetlig samfunnsdokumentasjon. I samme St.meld. ble Riksarkivaren bedt om å «utarbeide en ny strategi for privatarkivarbeidet i Noreg». Et av tiltakene som ble iverksatt var etableringen av SAMDOK, et paraplyprosjekt hvor underprosjekter skulle bidra til å styrke arbeidet med få en samlet samfunnsdokumentasjon. For privatarkivfeltet ble det bl.a. gjennomført prosjekter som skulle utvikle og samle erfaringer med utarbeiding av bevaringsplaner for privatarkiv, for å lage en felles metodikk for dette. To av disse prosjektene var utvikling av bevaringsplaner for privatarkiv i Aust-Agder[2] og i Nordland.[3] I rapporten "En helhetlig samfunnshukommelse"[4] sammenfattes de metodiske resultatene fra de to bevaringsplanene og gir en oversikt over hjelpemidler som ble utviklet for å støtte arbeid med å lage arkivplaner.

Først i dette kapitlet vil metoden for å lage bevaringsplaner som ble utviklet under SAMDOK-prosjektet oppsummeres. Bevaringsplaner utviklet etter den metoden vil ikke nødvendigvis si noe om digitalt skapte arkiv, men legger grunnlaget for å planlegge bevarings- og innsamlingsarbeid.

I siste del av kapitlet beskrives noen utfordringer og muligheter det kan være nyttig og nødvendig å ta hensyn til i forbindelse med planlegging av bevarings- og innsamlingsarbeid av digitalt skapt privatarkiv. Dette for å sikre at verdifull digitalt skapt samfunnsdokumentasjon eller annen viktig dokumentasjon ikke går tapt, og å utnytte noen av de mulighetene som er iboende i det digitale formatet.

2.1 Formål med bevaringsplan

Det grunnleggende formålet med en bevaringsplan er å sikre bevaring av et best mulig utvalg av privatarkiv, som gjenspeiler samfunnet og samfunnsutviklingen frem til nåtid. En bevaringsplan fungerer på to

[1] St.meld. 7 (2012-2013):
https://www.regjeringen.no/no/dokumenter/meld-st-7-20122013/id707323/
[2] BIPA-prosjektet «Bevarings- og innsamlingsplan for privatarkiv fra Aust Agder»:
https://www.kubenarendal.no/media/138938/Bevarings-og-innsamlingsplan-for-privatarkiv-fra-Aust-Agder-Rapport.pdf
[3] Bevaringsplan for privatarkiv i Nordland fylke:
http://www.arkivinordland.no/forsiden/aktuelt/bevaringsplanen-for-privatarkiv-i-nordland-fylke-er-ferdig.1001501.aspx
[4] Privatarkivutredingen:
https://samdok.com/2015/01/29/en-helhetlig-samfunnshukommelse

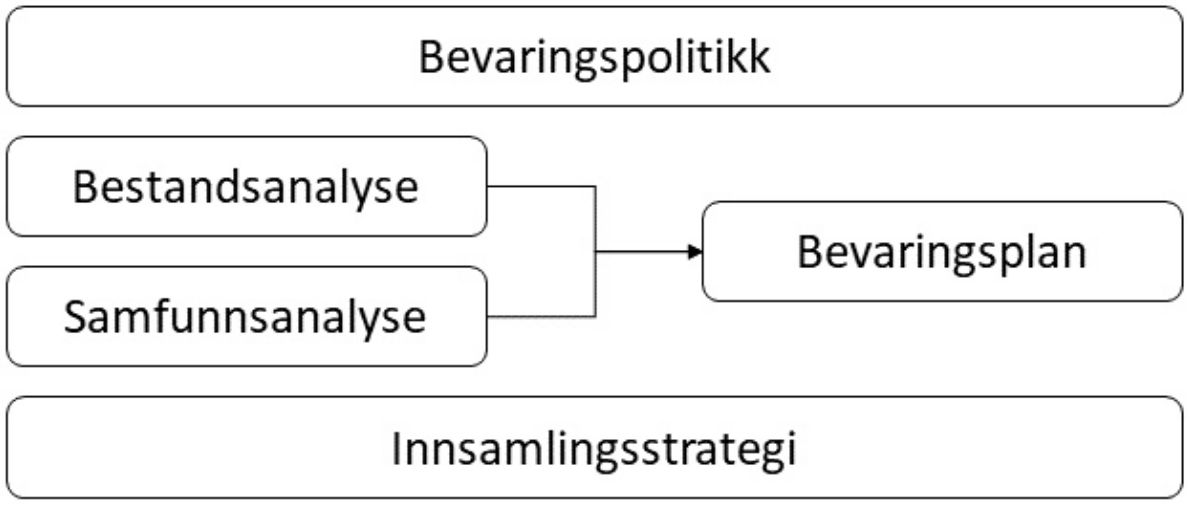
nivå: Et makronivå hvor bevaringsplaner fra forskjellige bevaringsinstitusjoner inngår som del av en samlet samfunnsdokumentasjon, og et mikronivå hvor den gir hovedlinjer og prinsipper for bevaring hos den enkelte bevaringsinstitusjon. Hovedlinjer og prinsipper for bevaring omtales som bevaringspolitikk. Den vil inneholde en beskrivelse over hvilket arkivmateriale en bevaringsinstitusjon tar sikte på å bevare.

På makronivå har Riksarkivaren gjennom "Retningslinjer for arbeid med privatarkiver" [5] beskrevet hvordan forholdet mellom bevaringsinstitusjoner skal reguleres for «å sikre at arbeidet med bevaring av privatarkiver utføres planmessig og systematisk», og for å styrke samarbeidet mellom bevaringsinstitusjoner. Riksarkivaren går videre med å beskrive bevaringspolitikk og bevaringsplaner som grunnlag for koordinering mellom bevaringsinstitusjoner. Koordinering og avtaler om arbeidsdeling mellom bevaringsinstitusjoner skal sikre bredde i bevaring av privatarkiv. Dette skal skje på, og rapporteres fra, fylkesnivå for bevaringsinstitusjoner med en regional tilknytning, mens bevaringsinstitusjoner med et nasjonalt arbeidsområde rapporterer direkte til Riksarkivaren.

På mikronivå er det opp til hver enkelt bevaringsinstitusjon å definere og beskrive sin bevaringspolitikk og utarbeide bevaringsplan i henhold til den.

2.2 Metode for utvikling av bevaringsplaner

SAMDOKs metode for utarbeidelse av bevaringsplan for privatarkiv, er en trinnvis tilnærming som består av tre deler: bestandsanalyse, samfunnsanalyse og bevaringsplan. I korte trekk skal resultatene fra de to analysene sammenstilles for å vise avvik i dekningsgraden i arkivbestanden, formulert i en bevaringsplan.



Figur 2.1: Bevaringsplan

Utgangspunktet for arbeidet med bevaringsplan er den enkelte bevaringsinstitusjons definerte bevaringspolitikk og eventuelle avgrensinger gjort mot andre bevaringsinstitusjoner. Bevaringspolitikken setter rammer for hva som skal inngå i bevaringsplanen, og er relativt stabil for den enkelte bevaringsinstitusjon. For bevaringsplaner er det annerledes. Under normale omstendigheter vil en bevaringsplan bli ansett som et styringsdokument for bevaringsinstitusjonen for en periode, men det vil gradvis oppstå behov for oppdatering eller revisjon i perioden den er satt til å fungere. Dersom planen bidrar til en planmessig og systematisk bevaring av samfunnsdokumentasjon, vil dette påvirke resultatene fra bestandsanalysen. I tillegg kan endringer

[5] Retningslinjer for arbeid med privatarkiver:
https://www.arkivverket.no/om-oss/vare-publikasjoner/riksarkivarens-rapporter-og-retningslinjer/_/attachment/download/52101cdf-72ec-4d16-91fe-bed6e2ec7a05:7742a1688d0dfec7874c111e1882cbc0c4aa084c/Retningslinjer%20for%20arbeidet%20med%20privatarkiver%20(Riksarkivaren%202005)%20Rapport%2011.pdf

i samfunnet over tid, og vår forståelse av endringene, påvirke resultatene fra samfunnsanalysen, som i sin tur kan gi grunnlag for å revidere bevaringsplanen.

2.2.1 Bestandsanalyse

I den først fasen, bestandsanalysen, kartlegges aktører og arkiver i arkivbestanden hos bevaringsinstitusjonen. Formålet er å gi en oversikt over hele arkivbestandens sammensetning i en bestandsoversikt. Informasjon om aktører og arkiv grupperes etter opplysninger eller metadata som beskriver dem. For å gjennomføre dette benyttes de metadata som arkivaren har tilført arkivene og aktørene i arkivkatalogen. Utvalgte metadata aggregeres for å kunne vise en sammenstilt fremstilling av både kvantitative og kvalitative forhold ved arkivbestanden.

Kvaliteten på og omfanget av relevante og målbare metadata i arkivkatalog er avgjørende for et godt og anvendbart resultat. Dette kan medføre behov for omfattende arbeid med katalogen [6].

For bevaringsinstitusjoner som bruker Arkivportalen, finnes det rapporter som kan benyttes for å hente data til analyse og videre prosessering. Der er det tilgjengelig rapporter over aggregerte data fra kriterier valgt av bruker. I arkivkatalogverktøyet ASTA krever aggregerte rapporter konsulentbistand. Verktøyene blir omtalt i Privatarkivutredningen [7]. Regneark eller database kan benyttes for prosessering av data fra arkivkatalogverktøy, eller fra data som er samlet inn i forbindelse med å lage bestandsoversikt.

Følgende informasjon om aktører kan være med på å dokumentere arkivbestanden:

- Kategorier som kan benyttes til å skille og gruppere aktører på et overordnet nivå.

  Bedrifter, organisasjoner og personer er mye brukte kategorier, men andre overordnede kategorier benyttet til å beskrive aktører kan benyttes.

- Underkategorier.

  For enkelte overordnede kategorier er det underkategorier som ytterligere kan skille og gruppere aktører og bidra til å gi et mer detaljert bilde av aktørene i arkivbestanden. For bedrifter og organisasjoner er næringskodene (SSB) mye brukt. Andre underkategorier kan også benyttes.

- Geografisk tilhørighet og virkeområde til den enkelte aktør.

- Datering av perioden en aktør har eksistert.

  Dette vil gjøre det mulig å plassere aktører på en tidsakse.

Følgende informasjon om arkiv kan være med på å dokumentere arkivbestanden:

- Datering av perioden innholdet i et arkiv dekker.

  Dette vil blant annet kunne bidra til å antyde noe om hvor komplett hvert enkelt arkiv er i forhold til perioden en eller flere aktører har fungert. Aggregert resultat over perioder kan vise om arkivbestanden har mangler eller er tilstrekkelig dokumentert.

- Ordningsgrad for hvert enkelt arkiv.

  Dette vil kunne indikere noe om den overordnede tilstanden til arkivbestanden.

- Kompletthet.

  Dette er et kriterium som forteller om i hvilken grad hvert enkelt arkiv i arkivbestanden anses å være komplett eller ikke. Det vil i aller høyeste grad være en subjektiv vurdering, og for de fleste bevaringsinstitusjoner vil dette være opplysninger som ikke allerede er beskrevet.

  Verdien av kriteriet i arbeidet med å lage en bevaringsplan er stor. Det har verdi for både det enkelte arkiv, men også for gruppering av arkiv innen kategorier. Utfordringen er blant annet å definere hva et komplett arkiv består av, og avsette de ressursene som kreves for å gjøre den faktiske vurderingen.

[6] BIPA-prosjektet, s. 7
[7] Privatarkivutredningen, s. 42

- Volum.

  Dette kan ha en indirekte verdi hvis man ser på forholdet mellom antallet arkiv innen en kategori og arkivenes volum. Ikke alle kategorier/arkiv kan ha forventet stor produksjon av dokumenter.

2.2.2 Samfunnsanalyse

Samfunnsanalysen beskriver og analyserer samfunnsutviklingen over tid, og hva som preger samfunnet av i dag. [8] Det er en selvstendig analyse i den forstand at den ikke skal følge gjeldende bevaringspraksis ved bevaringsinstitusjonen, eller preges av resultatene fra bestandsanalysen. Dette gjøres for å forhindre overrepresentasjon av deler av samfunnsdokumentasjonen, og for å øke sannsynligheten for en tilstrekkelig bredde i denne. Den må dreies mot de deler av samfunnet og den historiske samfunnsutviklingen som er relevant for bevaringsinstitusjonen og er inkludert i dens egen bevaringspolitikk.

Samfunnsanalysen bør dekke et tidsrom hvor det er sannsynlig at det kan ha blitt skapt arkiv med relevans for bevaringsinstitusjonen. Tidsrommet kan deles opp i epoker for å ivareta endringer i samfunnsutviklingen som kan påvirke hva som bør dokumenteres fra de enkelte epokene. Dersom man allerede er kjent med større mangler i arkivbestanden, kan enkelte perioder eller epoker i samfunnsanalysen vektlegges. [9]

Utarbeidelse av samfunnsanalyse krever god kunnskap om de utviklingstrekk, trender, endringer og sammensetninger som har preget samfunn og samfunnsutviklingen i det tidsrommet som skal dekkes.

2.2.3 Bevaringsplan

Resultat fra analysene av samfunnsutviklingen og arkivbestanden vurderes opp mot hverandre, for å vise dekningsgraden i arkivbestanden hos bevaringsinstitusjonen. Dekningsgrad betegner da i hvor stor grad et gitt relevant trekk i den historiske samfunnsutviklingen er representert i arkivbestanden.

I BIPA-prosjektet angis dekningsgraden for hver enkelt underkategori i bestandsanalysen for hver enkelt av de aktuelle periodene. Dekningsgraden deles der inn i fire kategorier:[10]

- Dokumentert, i betydningen en viss grad av dokumentasjon finnes.
- Svakt dokumentert.
- Ikke dokumentert.
- Uten dokumentasjonsgrunnlag, i betydningen marginalisert eller ikke forekommende kategori.

Videre gjøres en samlet vurdering av dekningsgrad. Denne er retningsgivende for det overordnede behovet for tiltak. Samfunnsanalysen tilfører en forståelse av hvor representativ og sentral underkategoriene er. Dette påvirker også den samlede vurderingen av hver enkelt kategori. I tillegg kan samfunnsanalysen avdekke kategorier/samfunnselementer som er underrepresentert eller ikke representert i bestanden.

Dekningsgraden innen de utvalgte kriterier eller kategorier, og eventuelle mangler i arkivbestanden som avdekkes, inngår i bevaringsplanen. Det er retningsgivende for videre arbeid med utvikling av bevarings- og innsamlingsstrategi.

2.3 Planlegge bevarings- og innsamlingsarbeid

Det videre arbeidet i forlengelsen av bevaringsplanen vil være individuelt for den enkelte bevaringsinstitusjon, men vil kunne bestå av identifikasjon av satsningsområder og prioriteringer i det videre bevarings- og innsamlingsarbeidet. I BIPA prosjektet medførte dette et behov for et skifte fra den tradisjonelle tilnærmingen, hvor man bevarer arkivene som blir avlevert, til en mer bevisst og aktiv innsamlingspolitikk.

[8] Privatarkivutredningen, s. 50
[9] BIPA-prosjektet, s. 17
[10] BIPA-prosjektet, s. 32

2.4 Trekk ved digitalt skapt privatarkiv

I det følgende vil vi presentere noen trekk ved digitalt skapt privatarkiv, som kan tas hensyn til ved planlegging av bevarings- og innsamlingsarbeidet. Noen av disse trekken vil sammenfalle med tilsvarende trekk ved arbeid med digitalt skap arkiv i offentlig sektor, mens andre er særegne for digitalt skapt privatarkiv.

Vi skal også se på noen virkemidler som kan brukes for å bidra til at verdifull samfunnsdokumentasjon blir bevart.

2.4.1 Plassering og lagringsmedium

Det kan være stor variasjonen i plassering eller lagring av dokumentasjon hos private aktører. Større organisasjoner og virksomheter har gjerne utstrakt bruk av fagsystemer for spesifikke oppgaver, gjerne omtalt som strukturerte data. Mindre organisasjoner, virksomheter og privatpersoner lagrer i større grad filer i mapper på disk, omtalt som ustrukturerte data. Dataene kan være lagret lokalt, eller eksternt hos en tjenesteleverandør.

Bruken av eksterne drifts- og lagringsløsninger er utbredt. Det er både sikkerhetsmessige, økonomiske og praktiske årsaker til dette. Det kan dreie seg om enkle og allment tilgjengelige lagringsløsninger for filer, e-posttjenester, driftsmiljø hvor bruker selv kan sette opp og forvalte ulike programvareløsninger, og lisens- og avtalebasert tilgang til spesifikke fagsystem. Variasjonen er svært stor.

I tilfeller hvor aktører kjøper driftstjenester, oppstår det et skille mellom de som kjenner innholdet i arkivene, og de som drifter og forvalter løsningene. I forbindelse med bevarings- og innsamlingsarbeid må man ha kontakt med begge parter. Tilgangen til data gis gjennom aktøren som eier eller forvalter informasjonsinnholdet. Denne kjenner også konteksten informasjonsinnholdet er skapt i, mens de som drifter løsninger kjenner de tekniske forholdene data forvaltes i. For løsninger hvor brukerene ikke selv forvalter og drifter, kan det være behov for å involvere tjenesteleverandør for å få gjennomført uttrekk av arkivdata.

I tilfeller hvor eksterne leverandører har lagret data på måter som gjøre uttrekk komplisert og tidkrevende, kan et alternativ være å gjennomføre en kost/nytte vurdering, dersom samme type dokumentasjon kan dekkes fra en annen kilde/aktør, hvor data er lettere tilgjengelig.

I bevarings- og innsamlingsarbeid hvor data befinner seg hos eksterne drifts- og lagringsleverandører, er man avhengig av å bruke metoder og verktøy som er fleksible nok til å kunne anvendes direkte hos leverandør, enten av leverandør selv eller sammen med bevaringsinstitusjonen, eller over nettverk. Metode og verktøy må være fleksibelt nok til å kunne håndtere store datamengder i uttrekkssituasjonen.

Når aktører benytter seg av eksterne tjenester hvor en tredjepart må involveres i forbindelse med bevarings- og innsamlingsarbeidet, øker kompleksiteten og tiden som trengs for å gjennomføre de fleste operasjoner, som en følge av antallet personer som er involvert og større grad av spesialisering (teknisk personell, brukere, administratorer og ledere).

Lokal oppbevaring av data på mindre lagringsenheter, som for eksempel harddisk i PC eller bærbare medier har vært et trekk ved digitalt skapt privatarkiv som har skilt det fra offentlige arkiver. Det er ikke bare benyttet av enkeltpersoner, men også mindre virksomheter, lag og organisasjoner. Det er vel rimelig å anta at bruken av eksterne lagringstjenester etter hvert har økt til erstatning for lokal lagring, men ikke erstattet lokal lagring fullt og helt.

Bruk av lokal lagring er svært sårbart, og faren for tap av viktig dokumentasjon er stor. Fysiske medier har begrenset levetid. De er sårbare for skadevare, tyveri, fysisk skade ved uhell, vannskade eller brann. For de fleste aktører er dette kjente problemstillinger, og noen iverksetter tiltak for å redusere farene. I tillegg til antivirusprogamvare blir ulike former for backup benyttet av mange. De fleste strategier er basert på manuelle prosesser og krever gode rutiner for å følge opp.

Oppsummert kan vi karakterisere data lagret eksternt som mindre utsatt for tap, men kan både være mer komplekse og tidkrevende å håndtere i forbindelse med bevarings- og innsamlingsarbeid enn data lagret lokalt. Data lagret lokalt er mer utsatt for tap og skade, men vil normalt være enklere å håndtere ved bevarings- og

innsamlingsarbeid. Begge disse perspektivene kan være faktorer som kan inngår i planlegging av bevarings- og innsamlingsarbeid.

2.4.2 Personavhengighet og eierskap

Tilgangen til data kan være personavhengig. Det er å forvente hos privatpersoner, men også hos mindre virksomheter og organisasjoner. Tilgangen til data reguleres gjennom enkeltpersoner som kjenner plassering, og eventuelle passord benyttet for å nå dataene. Større virksomheter og organisasjoner vil normalt være innrettet slik at tilgangen ikke skal være personavhengig. Personavhengighet gjør data sårbar som mulig samfunnsdokumentasjon.

Det kan også oppstå situasjoner hvor data blir eierløse. Det vil si når de som eier eller forvalter informasjonsinnholdet ikke lenger kan finnes, eller ikke lenger benytter eller betaler for eksterne lagrings- og driftstjenester. Dersom data blir eierløse, ved eksempelvis konkurser, nedleggelser, endringer i virksomheter og organisasjoner eller ved død, vil de i mange tilfeller raskt bli utilgjengelig og gå tapt. Eierløse data oppbevart hos eksterne lagrings- eller tjenesteytere vil ha en begrenset levetid før de slettes, enten som en konsekvens av manglende betaling for tjeneste eller når det ikke har vært bevegelser i bruken av et materiale (f.eks. e-post).

Personavhengighet og eierskap bør være en av faktorene som inngår i planlegging av bevarings- og innsamlingsarbeid.

2.4.3 Arkivskapere

Forhold ved arkivskapere og ansvaret disse har for et gitt materiale er også noe det kan være behov for å vurdere i forbindelse med planlegging av bevarings- og innsamlingsarbeidet. Kompetanse og kunnskap om teknologien som bærer og tilgjengeliggjør data vil variere. Dette kan påvirke deres mulighet til å gjøre veloverveide og gode valg ved håndtering av digitalt skapt materiale.

Situasjoner som kan påvirke valgene som tas er f.eks. når en bruker ikke selv kan nå innholdet som en konsekvens av manglende programvare, eller når vedkommende ikke har rett kompetanse til å kunne hente ut meningsfylt informasjon. Kostnader ved lagring av data over tid kan også påvirke valg. Arkivskaper kan også utføre forvaltningstiltak som endrer eller forringer informasjonsinnholdet i data sett fra et dokumentasjonssynspunkt.

2.5 Virkemidler

Det er flere virkemidler en bevaringsinstitusjon kan benytte seg av for å bidra til å motvirke eller forenkle utfordringer knyttet til bevarings- og innsamlingsarbeidet av digitalt skapt privatarkiv.

2.5.1 Tidlig kontakt

Tidlig kontakt innebærer å etablere kontakt med en arkivskaper mens den fremdeles er aktiv. Dette er en naturlig konsekvens av oppfølging av funn i bevaringsplan og er omtalt både i metodebeskrivelsen i "Privatarkivutredingen" [11] og i flere publiserte bevaringsplaner som del av det videre arbeidet [12] [13]. Det omtales gjerne som proaktiv innsamlingspolitikk.

Tidlig kontakt vil kunne bidra til en planmessig, systematisk og målrettet innsamling av samfunnsdokumentasjon, og en god utnyttelse av tilgjengelige ressurser hos den enkelt bevaringsinstitusjon.

[11] Privatarkivutredningen, s. 55
[12] BIPA-prosjektet, s. 43
[13] Arkivverktes bevaringsplan for privatarkiv, s. 1

Tidlig kontakt gir mulighet til å identifisere og kartlegge arkivmateriale for å kunne avdekke eventuelle utfordringer. Det er viktig å danne relasjoner til aktører som kan bistå med å finne digitale kilders plassering, gi tilgang til data og kartlegge hvilke tekniske forutsetninger som skal til for å inkludere dette i bevarings- og innsamlingsarbeid. Det kan bidra til å avdekke tekniske problemstillinger og hull i kompetansen hos bevaringsinstitusjonen, slik at det kan iverksettes nødvendig tiltak før bevaringsarbeidet starter.

Tidlig kontakt med arkivskaper åpner for å kartlegge og vurdere kildene før de eventuelt tas med som del av bevarings- og innsamlingsarbeidet. Det kan være deler av det digitale materialet som ikke er relevant eller som finnes i offentlige arkiver. Med tidlig kontakt hos flere aktuelle arkivskapere innen et felt hvor bevaringsplanen har avdekket mangler i dekningsgrad, vil man kunne gjøre en samlet vurdering og prioritering for å unngå unødvendig bruk av ressurser.

Innsamling av kontekstdokumentasjon direkte fra kilden vil kunne forenkle innsamling og potensielt øke kvaliteten på informasjon i aktør- og arkivbeskrivelser. Tidlig kontakt kan også gjøre det enklere å kartlegge relasjoner og nettverk mellom aktører, ikke bare som del av aktørbeskrivelse, men også for å utvide antallet mulige kilder innen enkelte felt.

2.5.2 Tillit

Hos enkelte arkivskapere vil det være behov for å bygge tillit før et arkiv kan overleveres til en bevaringsinstitusjon. Det kan være uklarheter knyttet til hvordan arkivet kan brukes, hvem som kan bruke det, eierskap og forvaltning av materialet når det befinner seg hos bevaringsinstitusjon. Også for arkivmateriale etter utsatte grupper kan det være avgjørende at man iverksetter tiltak utover ordinære personvernhensyn for å sikre at det er mulig å bevare dokumentasjon. Det kan også være forretningsmessige hensyn som må tas i forbindelse med bevaring av privatarkiv.

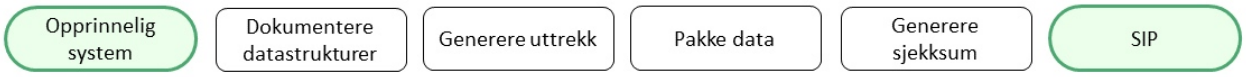
Bygge tillit kan gjøres med informasjon om hvordan arkivskaperens interesser blir ivaretatt når arkivmaterialet blir innlemmet som del av arkivbestanden. I noen tilfeller kan det være nødvendig å bygge personlige relasjoner for å få tilstrekkelig tillit hos arkivskapere.

2.5.3 Tidlig uttrekk

Det er også mulig å planlegge for å gjennomføre bevarings- og innsamlingsarbeid før arkivmaterialet går ut av aktiv bruk hos aktør. I enkelte tilfeller kan det være nødvendig å gjøre dette for å sikre at verdifull dokumentasjon ikke går tapt som en følge av noen av utfordringene nevnt tidligere. Her er det ikke noen regelverk som stiller krav til når arkiv bevares.

Kapittel 3

Uttrekk fra opprinnelig system



Figur 3.1: Prosesser i et uttrekk fra opprinnelig system til ferdig SIP

Vi skal nå se hvordan man kan hente ut data fra opprinnelig system eller plassering, for å lage det OAIS kaller en SIP (submission information package).

I dette og de påfølgende kapitlene "Normalisering", "Dokumentasjon", "Tilgjengeliggjøring" og "Forvaltning av digitalt skapt privatarkiv" bruker vi et forenklet flytdiagram for å illustrere innputt, utputt og de mest sentrale handlingene som blir utført i den delen av prosessen som beskrives i kapitlet.

Når man skal hente ut data fra opprinnelig system og lagringssted, er det en fordel at dataene beholdes mest mulig på den form og struktur som de ble skapt. Dette er viktig både for integritet og autentisitet, men også for å være sikker på at man får med seg alle viktige data. All omforming av data medfører risiko for at man kan miste noe på veien.

Resultatet av uttrekket skal altså være en pakket fil (SIP) med data som er minst mulig endret, og hvor mest mulig metadata er beholdt.

3.1 Dokumenter

Man bør hente ut dokumentene på det formatet de ligger, og ikke forsøke å konvertere dem før man henter dem ut. Konvertering er en del av normaliseringen (jf. neste kapittel). Filene bør hentes ut i den mappestrukturen de ligger. Mappestrukturen kan gi viktig kontekstinformasjon for dokumentene.

Det er også viktig å få med alle metadata som ikke ligger i selve filen, men på filsystemet. Der ligger informasjon om når filen ble opprettet og av hvem, når den ble endret og av hvem, og eventuelt andre metadata.

Se under "Verktøy" hvordan Bergen byarkiv håndterer dette.

3.2 Databaser

For databaser er det viktig å hente ut alle data fra databasen på den struktur databasen opprinnelig hadde. Det vil si at vi henter ut alle rader fra alle tabeller fra opprinnelig database, og all info om hvordan databasen er bygd opp og henger sammen.

For offentlige arkiver har det vært en utbredt praksis enkelte steder å konvertere databasen til en helt annen struktur i uttrekket, og luke bort store mengder data. Men dette kan i verste fall føre til datatap. Det er vanskelig alltid å vite hvilken informasjon i en database som er viktig å bevare for ettertiden. Det kan også forekomme feil i relasjoner i opprinnelig base som fører til at mye data ikke kommer med på uttrekket. Derfor bør hele basen tas med i uttrekket, og så kan man heller tilgjengeliggjøre den i en annen struktur. I OAIS-terminologi så betyr det da at SIP bør være opprinnelig database, mens DIP kan være en omstrukturering.

Det verktøyet som de fleste arkivinstitusjoner bruker for å ta uttrekk fra relasjonsdatabaser, er SIARD. Det henter ut informasjon om strukturen til databasen og legger i en xml-fil. Tilsvarende blir selve dataene i databasen lagret i én xml-fil per tabell. SIARD kan da også brukes til å laste data opp i en ny database igjen.

Utenfor arkivmiljøer er det istedenfor SIARD vanlig å bruke det standardiserte språket som allerede finnes for relasjonsdatabaser, nemlig SQL, for database-migrasjon. Ved å bruke SQL direkte (det blir også brukt internt i SIARD) kan man få ut nøyaktig SQL brukt for å opprette en database, og slik ta vare på absolutt all info, som triggere, indekser, views osv. Det blir også veldig enkelt å laste opp igjen data i en database, og man behøver ikke spesialiserte verktøy som SIARD til det.

Bergen byarkiv har valgt å generere SQL som følger ISO-standarden. SQL ble nemlig en ISO-standard allerede i 1987. Selv om det finnes en del unntak fra denne standarden i ulike databaser, og ikke alle databaser oppfyller like mange deler av standarden, er nå støtten såpass bra, at det skal små endringer til for å tilpasse denne standarden til ulike databasemotorer. Støtten blir også stadig bedre i alle de største databasemotorene, år for år. Vi har dermed et format vi vet vil være håndterbart i veldig mange år framover, og som ikke er avhengig av at spesifikk programvare for arkivsektoren blir vedlikeholdt.

Selve dataene kan man ta ut som SQL-inserts (altså setninger som sier hva som skal settes inn i hver tabell), eller på et tegnseparert format. Bergen byarkiv har valgt TSV-formatet, altså en tekstfil med én post per linje, hvor kolonnene er atskilt med tabulator. En nærmere teknisk diskusjon omkring dette kan leses under Verktøy.

3.3 Pakking av data

Etter å ha generert selve uttrekket, bør man pakke det og generere en sjekksum på den pakkede filen. Denne sjekksummen oppbevares på en slik måte at man kan kontrollere at ingenting er endret i SIP-filen.

For mer detaljert og teknisk diskusjon om uttrekk og ulike formater, se Verktøy.

Kapittel 4

Normalisering

Figur 4.1: Prosesser i normaliseringsfasen, hvor man begynner å bearbeide opprinnelig SIP til AIP

Nå skal vi se hvordan uttrekket kan normaliseres for å inngå i en AIP (archival information package).

Med normalisering av data menes at data konverteres til formater som er mest mulig programuavhengige og anvendbare for framtiden.

Se under Verktøy for mer detaljert og teknisk diskusjon rundt normalisering, og om mulige verktøy for å gjøre jobben.

4.1 Dokumentasjon av endringer

Alle de endringer som gjøres blir dokumentert automatisk av programvaren vi bruker. Og det bør alltid foretas en sjekk av at opprinnelig sjekksum på uttrekket fremdeles stemmer før man går i gang med normaliseringen.

Før normalisering kan man foreta en grov arkivbegrensning, dvs. å fjerne data som åpenbart ikke er arkivverdig, for å slippe å drive normalisering på disse. Dette dreier seg da om å fjerne hele tabeller eller mapper. Endringer her bør også fanges opp automatisk av programvaren og dokumenteres.

4.2 Dokumenter

Normalisering av elektroniske dokumenter vil si at vi konverterer dem til arkivformat. Dette vil oftest være PDF/A, men det finnes også andre aktuelle arkivformater. En bør også tenke på at man bør ha et format som egner seg best mulig for tilgjengeliggjøring. Derfor konverterer Bergen byarkiv dokumenter i størst mulig grad til formater som lar seg åpne i en nettleser (jf. under Tilgjengeliggjøring). Formater som lar seg åpne i nettlesere, egner seg ofte godt til langtidslagring også, da det er veldig utbredte formater, med god verktøystøtte.

Dokumenter som av ulike årsaker ikke lar seg konvertere til arkivformat, kan man beholde i sitt opprinnelige format. Da har man mulighet til eventuelt å konvertere dem senere hvis det dukker opp nye konverteringsløsninger. Man kan ellers lese dem i dedikert programvare, eller hente ut informasjon på annen måte.

Dokumenter hvor en konvertering kan føre til datatap, bør man også beholde i sitt opprinnelige format, i tillegg til arkivformat. Det gjelder spesielt Excel-filer, hvor formler m.m. forsvinner ved konvertering til PDF.

Det er også mulig å beholde opprinnelig format for alle dokumenter hvis man ønsker det. Da forblir de opprinnelige formatene i SIP-en, mens de konverterte legges i AIP. Om man ønsker å gjøre dette avhenger stort sett av om man tar seg råd til å oppbevare de opprinnelige dokumentene i tillegg til arkivformatene. Lagringsplass kan være veldig dyrt, hvis man har store mengder med dokumenter.

4.3 Databaser

For databaseuttrekk kan man tilsvarende gjøre normalisering av tabelldata slik at disse er helt databasemotoruavhengige.

Dersom basen inneholder dokumenter lagret i selve databasen, trekkes disse ut og lagres på disk. Da må man legge inn en referanse til dokumentene i tabellen de ble fjernet fra.

Hvis man lagrer data i tegnseparerte filer, må man normalisere innholdet i forhold til det formatet som er valgt. Ved Bergen byarkiv bruker vi TSV (tabulator-separerte-verdier), og da må vi fjerne tabulator fra teksten, og erstatte med mellomrom.

Navn på tabeller og kolonner normaliserer vi til små bokstaver (for å virke likt på tvers av alle databasemotorer), og endrer litt på navn som kan være beskyttet (dvs. ha spesiell betydning) i enkelte databaser. Da er vi sikre på å få lastet dem inn i ny database uten at noe feiler.

Og så normaliserer vi SQL for å opprette en database til mest mulig å samsvare med ISO-standarden. Vi normaliserer disse så mye som mulig, slik at de eneste tilfellene av inkompatibiliteter som gjenstår ift. enkelte databasemotorer, er noen få datatyper (eksempelvis timestamp i mssql, som er noe helt annet enn timestamp i alle andre databaser).

4.4 Samling av data i arkivpakke (AIP)

Resultatet av normaliseringen, dvs. alle konverterte dokumentfiler, normaliserte TSV-filer fra databaser, ISO SQL som beskriver databasen, samt beskrivelse av all normalisering som er gjort, samles i en mappe som skal danne grunnlaget for arkivpakken.

Men før denne pakken kan lagres i DSM, må man dokumentere uttrekket godt. Og det skal vi se på i neste kapittel.

Kapittel 5

Dokumentasjon

Figur 5.1: Prosesser for å dokumentere et uttrekk og ferdigstille AIP med DIP

Dokumentasjonen skal sikre at alle data som legges i arkivpakken forstås innenfor den konteksten de ble skapt. Man må dokumentere hva data betegner, koblingen mellom ulike data, og ansvar og virkeområde for arkivskaperen som skapte dataene.

Spesielt for vår metodikk er at vi bruker innsynsløsningene som genereres, som del av dokumentasjonen. Ettersom vi legger skript for å generere innsynsløsning inn i arkivpakken, vil disse skriptene være dokumentasjon av hvordan dataene skal forstås.

5.1 Overordnet dokumentasjon

I katalogen registreres opplysninger om arkivet eller arkivenheten som det er tatt uttrekk av. Dette vil typisk være registrering på arkiv- og serienivå. Det skal også registreres informasjon om arkivskaperen. Disse opplysningene legges så i arkivpakken. Den greieste måten å gjøre dette på, er å ta en eksport av opplysningene fra katalogbasen, gjerne i form av EAD- og EAC-CPF-filer.

5.2 Dokumenter/filer

Som nevnt ovenfor er alle dokumenter gjennomgått av programvare som henter ut all metadata om dokumentene, og alle dokumentene er normalisert (dvs. konvertert til arkivformat). Alt dette er dokumentert i en fil som har oversikt over alle dokumentene og resultatet av konverteringen.

Denne dokumentoversikten kan så importeres til en database for å produsere en DIP. Her tas med filnavn, evt. tittel, fildato og relativ filbane til den konverterte filen. Man kan også velge å registrere alle mappene i denne tabellen, slik at man kan navigere seg rundt i filstrukturen litt på samme vis som i et filsystem.

Hvis man ønsker det, kan man også gå gjennom filstrukturen manuelt, og legge inn en kolonne med beskrivelse av hva de ulike mappene eller filene inneholder. Om dette er hensiktsmessig å gjøre, avhenger naturligvis av hvor stor filstrukturen er og hvor mye arbeid det innbefatter.

Selve arkivet beskrives i katalogprogramvaren (f.eks. Asta), og her kan også seriene beskrives. Men så stopper som regel registreringen i katalogprogramvaren, og materialet beskrives videre i tabellen med filoversikt som genereres ved konvertering.

Sammen med katalogens overordnede beskrivelse av hva arkivet inneholder, vil filen med oversikt over dokumentene (og evt. mappene) utgjøre dokumentasjonen for uttrekk som kun består av filer. Den vil altså utgjøre katalogen til dette uttrekket. For på samme måten som man lager katalog for papirmateriale ved å beskrive mapper og eventuelt dokumenter, så vil tabellen med oversikt over elektroniske mapper og dokumenter beskrive disse filene i katalogen for det elektroniske materialet.

Vi ser ingen grunn til å importere denne beskrivelsen av filer inn i katalog-databasen (Asta), da vi uansett ikke tilgjengeliggjør filene i Asta. Vi unngår da fordobling av katalogdata ved å kun ha filbeskrivelsene i innsynsløsningen.

Det er da viktig å sørge for at alt av katalogdata man ellers ville hatt i katalogsystemet, legges inn i tabellen med oversikt over filer. Det man da særlig må passe på å få med, er informasjon om noen av filene eller mappene skal klausuleres. Disse opplysningene er jo vanskelig å få inn i hovedkatalogen (Asta), da hver enkelt mappe og fil ikke beskrives der.

Tabellen over filer er beskrevet i en fil (vi bruker tsv-format) som lagres som del av arkivpakken. Men den må samtidig tilgjengeliggjøres i en innsynsbase, og ses i sammenheng med øvrige katalog-data (jf. kapitlet "Tilgjengeliggjøring")

5.3 Databaser

5.3.1 Kartlegge opprinnelig database

Dersom det følger med en database i uttrekket, må denne databasen kartlegges grundig for å finne ut av sammenhenger og hva de ulike tabellene og kolonnene betegner, og hva som er aktuelt å ta med i en innsynsdatabase.

Det er en fordel om det kan skaffes til veie dokumentasjon av databasestrukturen. Leverandørene kan ofte sitte på slik dokumentasjon.

Uansett er det nesten helt nødvendig å ha tilgang til opprinnelig applikasjon som dataene er produsert i, for å kunne dokumentere dataene på en god måte. Har man tilgang til å se på dataene i den opprinnelige applikasjonen, vil man mye lettere finne ut av hva tabeller og kolonner betegner, og finne relasjonene i databasen.

Det er fremmednøklene i en database som viser hvordan dataene henger sammen, dvs. hvordan verdien av en kolonne i en tabell henviser til en post i en annen tabell. Men fremmednøkler kan ofte mangle i databaser, og da kan det være veldig vanskelig å finne sammenhengene uten tilgang til selve applikasjonen. Har man tilgang til applikasjonen og registreringsbildene der, ser man ofte fra nedtrekkslister hvilken tabell den underliggende kolonnen viser til.

I brukergrensesnittet ser man også lett hva som er de mest sentrale dataene i basen, og kan slik lettere identifisere de viktigste tabellene. Ledetekstene som er brukt i applikasjonen viser som regel hva kolonner i basen egentlig betegner. For kolonnene i databasen kan ofte ha kryptiske navn, og det gjør det vanskelig å finne ut av hva de er brukt til, hvis man ikke har tilgang til selve applikasjonen.

Det er også nyttig å kunne høre med dem som har brukt systemet om hva ulik info i databasen betyr. Ofte kan det være brukt koder som ikke er dokumentert noen steder i databasen. Hvis det er det, kan man dokumentere disse kodene ved at man enten oppretter en oppslagstabell i innsynsdatabasen (jf. nedenfor), eller at man erstatter kodene med det de faktisk betegner i innsynsbasen. Tilsvarende kan felt ha blitt brukt til noe annet enn hva systemleverandør la opp til (og navnga feltet som).

Det er i tillegg en fordel å ha verktøy som kan hjelpe til med å analysere en database. Det er mye informasjon som kan hentes ut direkte fra databasen, og som kan bidra til å bedre få oversikt over hvilke deler av basen som har vært i brukt og hvordan den henger sammen. Jf. dokumentet om verktøy.

5.3.2 Mapping av data til DIP

Mens kartleggingen er viktig for å finne ut hvordan basen henger sammen, og hva ulike tabeller og kolonner betegner, så er det mappingen til en innsynsløsning - en DIP i OAIS-terminologi - som vil være den egentlige dokumentasjonen av løsningen.

En mapping betyr at man oppretter en ny database (den databasen man vil bruke i innsynsløsningen), og dokumenterer hvordan innsynsbasen henter info fra opprinnelig database. Det blir altså en oversikt (mapping) av hvilke kolonner fra opprinnelig base som skal til hvilke kolonner/attributter i innsynsbasen.

En mapping kan altså i mange tilfeller erstatte en kartlegging av opprinnelig database. Det er ikke alltid hensiktsmessig å dokumentere absolutt alle tabeller og felter i en database. Slik blir en mapping en dokumentasjon av de tabellene og feltene som vi ønsker å ta med i en innsynsløsning.

En innsynsversjon vil ha med kun de mest relevante dataene. I denne prosessen foregår dermed også en bortskrelling av informasjon som ikke er særlig interessant. En database inneholder jo mange tabeller og kolonner som ikke er verdt å ta vare på. Det kan være pga. at de inneholder teknisk info som er nødvendig kun i forbindelse med drifting av opprinnelig system, det kan være tabeller eller kolonner som ikke (eller nesten ikke) har vært i bruk, eller det kan være info av kortvarig interesse, som man ikke behøver i en innsynsløsning.

Hvilken type mapping man gjør, har naturligvis sammenheng med hvilken innsynsløsning man bruker. Ved Bergen Byarkiv legger vi innsynsdataene inn i relasjonsdatabaser. Det er også relasjonsdatabaser som er brukt desidert mest til å lagre data hos arkivskaperne. Dermed får vi en ganske enkel måte å mappe på, ved at vi skriver SQL som mapper fra opprinnelig relasjonsdatabase til innsynsdatabasen.

Vi lager tabellnavn og kolonnenavn som er mest mulig beskrivende, slik at de også kan brukes som ledetekster i innsynsløsningen. Vi oppretter fremmednøkler som viser relasjonene, og vi oppretter indekser som gjør at oppslag går raskere, samtidig som vi navngir indeksene på en slik måte at de kan brukes til å dokumentere hvordan innsynsløsningen ser ut og skal fungere.

Vi bruker indekser til å angi standard sortering, hvilke kolonner som skal vises i trefflisten for hver enkelt tabell, samt hvilke kolonner som betegner en filbane. Slik får vi en selvdokumenterende DIP, som vi automatisk kan opprette en innsynsløsning for.

En slik DIP som kun består av sql-setninger for å opprette en database, er egentlig ganske uavhengig av spesialiserte verktøy for å vises fram seinere. Den er ikke basert på et spesielt skjema (i f.eks. xml eller json) for å beskrive hvordan data skal presenteres. Alt ligger i selve basen. Man kan altså i stor grad finne fram i basen bare ved å søke direkte i databasen med et SQL-verktøy. Men man har også mulighet til å få basen opp i innsynsløsninger som støtter de få reglene som er brukt for å bestemme hvordan data skal vises fram. Hos Bergen byarkiv har vi implementert disse reglene i en open source programvare som heter URD (jf. under "Verktøy"). Den blir da en slags referanseimplementasjon av hvordan man lager automatisk innsynsløsning basert på databasestruktur.

SQL-setningene som brukes for å opprette innsynsbasen og fylle den med data, lagres så i arkivpakken, som en dokumentasjon av uttrekket.

5.4 Arkvivbegresning

Når man har kartlagt filer og databaser, har man fått mer kunnskap om det finnes mer materiale her som kan arkivbegrenses. Vi så i kapitlet om normalisering at man kan foreta arkivbegrensning på materiale som åpenbart skal arkivbegrenses, for å slippe å normalisere dette materialet. Men på det tidspunktet har

man ganske mangelfull oversikt over materiale. Derfor kan man foreta en ny arkivbegrensning etter å ha gjennomgått og dokumentert materialet.

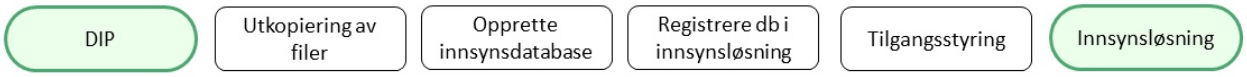
All arkivbegrensning skal også dokumenteres. Man må dokumentere at man har foretatt begrensningen, og gi en begrunnelse på hvorfor. Det er en fordel om man har verktøy som kan sørge for at dette blir gjort på en god måte. Vi har begrenset støtte for dette i verktøyene vi bruker i dag, men dette kommer etterhvert (jf. kapittel Verktøy).

Selv om man fjerner noe fra en arkivpakke, kan man godt beholde det i SIP-en. Slik kan man ta vare på det uten å måtte forholde seg til det seinere.

Når all dokumentasjon er på plass, så kan man ferdigstille arkivpakken, og legge den inn i DSM, jf. kapittel om forvaltning.

Kapittel 6

Tilgjengeliggjøring



Figur 6.1: Prosesser for å produsere en innsynsløsning fra en DIP

Når vi har en DIP liggende i arkivpakken, er det en veldig enkel sak å tilgjengeliggjøre dataene. Man kan bare kopiere ut alle filene man behøver, dvs. dokumentfiler, filer med tabelldata, og sql-skript for å opprette innsynstabeller. Så kjøres skriptene for å generere innsynsbasen, og basen registreres i innsynsløsningen. Deretter opprettes de tilgangene som behøves, og så er innsynsløsningen ferdig.

6.1 Viktigheten av rask tilgjengeliggjøring

Opprettelse av innsynsdatabase skjer egentlig i dokumentasjonsprosessen (jf. over), ettersom vi betrakter innsynsbasen som en del av dokumentasjonen. Tilgjengeliggjøring er følgelig en prosess som skjer i forbindelse med generering av arkivpakken. Og dette er viktig, ettersom det er mye enklere å få tilbakemelding fra arkivskaper om forhold som bør rettes opp i innsynsløsningen hvis man gjør den tilgjengelig for arkivskaper umiddelbart etter deponeringen. Dermed sikres best mulig dokumentasjon av uttrekket.

Hvis man tar imot et uttrekk, forsøker å dokumentere det, og bare legger det i en arkivpakke uten å opprette en innsynsløsning, vil man kunne få seg en overraskelse når man flere år etterpå forsøker å lage en innsynsløsning. Da er det ikke sikkert man har dokumentert uttrekket godt nok til å vite hvordan data skal vises fram.

For privatarkiver uten database, er det tabellen med filoversikten som tilgjengeliggjøres (jf. kapittel "Dokumentasjon"), og det er da de tilleggsattributtene som tas med der, med beskrivelse og eventuelt klausulering, som kan kvalitetssikres av arkivskaper.

6.2 Generiske innsynsløsninger er best

Innsynsløsninger bør være enklest mulig å opprette. Man bør bruke generisk programvare som kan vise fram alle mulige databaser. Det er for mye jobb å kode innsynsløsning for hver enkelt database, og det gjør migrering til nye løsninger veldig arbeidskrevende.

Det er også ved at vi bruker en generisk løsning at Bergen byarkiv kan lagre DIP-en i arkivpakken.

Med den selvdokumenterende DIP-en vi har i arkivpakken, er det en forholdsvis enkel sak å lage skript som oppretter en innsynsløsning for generiske systemer for framvisning. Ellers kan den også vises fram direkte i

URD, eller evt. andre systemer som følger samme standard for selvdokumenterende databaser (Jf. kapittel "Verktøy").

Ved Bergen byarkiv har vi en annen innsynsløsning enn den selvdokumenterende innsynsløsningen vi legger i arkivpakken. Det vil si at vi selv må kode hvordan databasen skal vises fram. Men da er det enkelt på et senere tidspunkt, når eksisterende innsynsløsning skal byttes ut, å få generert ny innsynsløsning basert på DIP-en i arkivpakken.

Ved opprettelse av innsynsløsning er det viktig å ta høyde for eventuelle behov for tilgangsstyring. Slike bør også gjøres mest mulig generisk, og slik at man kan styre tilgangen ganske detaljert for hver enkelt database. Se beskrivelsen av URD under "Verktøy" hvordan dette er tenkt implementert generisk der.

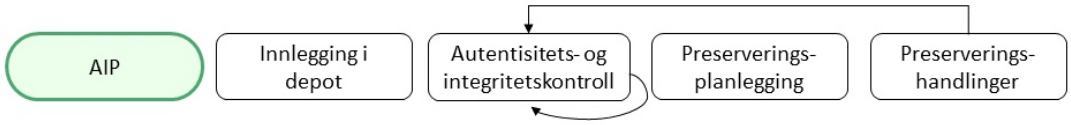
6.3 Integrasjon med katalog

Det er viktig å huske på at innsynsdatabasene er en del av katalogen. Derfor er det også ønskelig at disse integreres med katalogdatabasen på best mulig måte. Ved framvisning av katalogen, bør man kunne se data fra innsynsdatabasene når man ser på serien som representerer det elektroniske uttrekket. Og ved søk i katalogdata, bør man også få treff på data som ligger i innsynstabellene. Arkivarer som betjener arkivene vil da kunne finne både arkiv og data i samme løsning, samt at det potensielt vil kunne fungere som selvbetjening for publikum.

I prosjektet for ny katalog ved Bergen byarkiv, ser vi på hvordan en slik katalog kan implementeres. Vi kommer også til å lage en demoversjon av en katalog som fungerer på denne måten. Vi håper det blir mulig å demonstrere dette i en eventuell videreføring av dette prosjektet.

Kapittel 7

Forvaltning av digitalt skapt privatarkiv



Figur 7.1: Prosesser for å legge en AIP i detpot og forvalte den

Et digitalt depot skal sikre at digitalt skapt arkiv blir ivaretatt slik at det er tilgjengelig i et langt tidsperspektiv. Dette er det fundamentet som bevaring av viktig samfunnsdokumentasjon bygger på. De primære funksjonene som inngår i et digitalt depot, er å legge til rette for at digitalt skapt arkiv kan mottas, garantere at materialet ikke går tapt eller at informasjonsinnholdet endres, sikre at materialet som forvaltes der vil være anvendelig både i nåtid og i fremtid, og legge til rette for at materialet kan bli gjort tilgjengelig for ulike brukere. I tillegg kommer organisering og styring av funksjonene som inngår i et digitalt depot. Disse funksjonene er beskrevet i referansemodellen OAIS - Open Archival Information System.[1]

Bevaringsinstitusjoner vil legge litt ulik forståelse i av hva som inngår i et digitalt depot, ut fra de behovene og den funksjonen de selv har. For eksempel vil det være forskjeller på om tilretteleggelse av data for langtids tilgjengelighet inngår i de funksjonene som et digitalt depot har, eller om et digitalt depot er en mer passiv mottaker av ferdig tilrettelagt digitalt skapt arkiv.

Bevaringsinstitusjoner som skal ta imot digitalt skapt privatarkiv, bør etablere et digitalt depot. Det er viktig å ha både god teknisk kompetanse og arkivkompetanse tilgjengelig når man etablerer og forvalter et digitalt depot. Dette gjelder ved etablering av digitalt depot i egen virksomhet eller i samarbeid med andre. Etablering av et digitalt depot, som også kan fungere som et utgangspunkt for videre utvikling, kan gjennomføres relativt raskt uten større investeringer.

I det følgende viser rapporten noen enkle verktøy som kan benyttes ved etablering og forvaltning av et digitalt depot, og noen prinsipper som ligger til grunn for verktøyene.

Videre vil rapporten vise hvordan et enkelt og lett tilgjengelig verktøy for å håndtere og overvåke digitalt skapt arkiv er tatt i bruk hos Bergen byarkiv, og hvordan dette inngår som del av den helhetlige metoden som er presentert i denne rapporten.

[1] OAIS: https://public.ccsds.org/pubs/650x0m2.pdf

7.1 Etablering og forvaltning av digitalt depot

Ved etablering og forvaltning av et digitalt depot er det viktig å underbygge tiltro til at innholdet blir ivaretatt på en sikker måte. OCLC (Online Computer Library Center) definerte et troverdig (trusted) digitalt depot som «(...) one whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future».[2] Sentralt i denne beskrivelsen av et troverdig digitalt depot, er forståelsen av at det er et kontinuerlig arbeid eller en prosess. Dette forutsetter oppfølging og videreutvikling av det digitale depotet.

Det er utviklet en rekke hjelpemidler i form av standarder og veiledere som kan brukes ved etablering og forvalting av et digitalt depot. Utgangspunktet for de fleste er OAIS-standarden, men det varierer i hvor stor grad de dekker hele eller kun deler av OAIS som referansemodell. De varierer også i kompleksitet, og hvor mye ressurser som kreves for å følge opp retningslinjene.

Her er tre eksempler:

Trusted Digital Repository (TDR) Checklist - ISO 16363,[3] [4] tidligere TRAC. TRD har sjekkliste over krav som bør oppfylles, samt hvilken dokumentasjon som kreves til oppfølging av disse. Til sammen gir TRD en vurdering av hele organisasjonens pålitelighet, forpliktelse og beredskap ved drift av et digitalt depot, organisatorisk inndelt i tre deler: infrastruktur, forvaltningen av data og teknologi, teknologisk infrastruktur og sikkerhet.

Digital Preservation Capability Maturity Model (DPCMM)[5] er knyttet til både OAIS og TRD. Det er et verktøy for å forberede og planlegge forbedringer i en organisasjons evne til å ivareta og drive et digitalt depot. Det digitale depotet måles i modenheten (Maturity) som deles inn i 5 nivå. Ved gjennomgang eller revisjon av organiseringen for det digitale depotet, plasseres det på ett av modenhetsnivåene, og viser hvilke tiltak som må iverksettes for å gå til neste nivå.

Levels of Digital Preservation[6] er rettet mot organisasjoner som ønsker å begynne eller forbedre sin evne til å ivareta digitalt skapt materiale. Denne er enklere enn de to foregående, og er primært sentrert rundt funksjoner rundt teknisk bevaring av data. Funksjonene er inndelt i lagring, integritet, kontroll, metadata og innhold. Modenheten innen hvert område plasserer dem i nivåer fra 1-4, kjenn innholdet, beskytt innholdet, overvåk innholdet og oppretthold innholdet.

Hjelpemidlene nevnt over kan brukes om hverandre. Bruken avhenger av hva man skal oppnå, etablering eller utvikling og forbedring av det digitale depotet.

7.2 Preserveringsplanlegging

En sentral oppgave i et digitalt depot er å sikre at materialet som forvaltes der er anvendelig og kan gjøres tilgjengelig når det er behov for det. I et digitalt depot utføres preserveringshandlinger for å ivareta dette. En sentral preserveringshandling er å sikre at filene som oppbevares i depotets lagringsløsning, kan åpnes med tilgjengelig teknologi, og at informasjonsinnholdet kan forstås slik det opprinnelig fremstod. Derfor forsøkes alle filene å konverteres til et standard arkivformat før de lagres i depotet.

Men det kan finnes filer som ikke kunne konverteres da de ble lagt i depoets lagringsløsning, pga. manglende programvare for å kunne konvertere. Digitale depot som oppbevarer privatarkiver, vil normalt ha en større mengde av slike filformater som en følge av mindre regulering og styring av danningsprosessen enn det som er tilfelle i arkiver etter offentlig virksomhet.

[2] OCLC: https://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf
[3] TAD: https://www.iso.org/standard/56510.html
[4] TRAC: https://public.ccsds.org/pubs/652x0m1.pdf
[5] DPCMM: https://www.statearchivists.org/resource-center/resource-library/digital-preservation-capability-maturity-model-dpcmm/
[6] LoDP: https://ndsa.org/publications/levels-of-digital-preservation/

Det vil være behov for å iverksette preserveringshandlinger når det kommer filformater med gode kvaliteter for langtidstilgang som kan erstatte originale filformater, eller når teknologien som benyttes for å åpne filer fases ut.

7.2.1 Filregister

For å holde oversikt over hvilke filformater som oppbevares i depotets lagringsløsning, bør det opprettes et register over alle filer og versjoner av dem, format og plasseringen til hver enkelt fil. Filregisteret inngår som et viktig verktøy i planlegging og gjennomføring av preserveringshandlinger for materiale lagret i depotet.

Gjennom registeret kan filer med samme format grupperes for å identifisere omfang og plassering ved planlegging eller gjennomføring av filkonvertering som nødvendig preserveringshandling. Registeret bør ikke bare være begrenset til filer direkte knyttet til arkivmateriale, men også inkludere dokumentasjonen som er lagt til i arkivpakkene. Disse må også forvaltes over tid.

7.2.2 Depotregister

For å holde oversikt over innholdet og tilstanden til alt materiale som forvaltes i et digitalt depot bør man opprette et register over alle AIPer i lagringsløsning, samt SIPer og andre sentrale datasett som oppbevares utenfor lagringsløsning. Innholdet som beskrives er av teknisk karakter for de enkelte AIPer, SIPer eller andre datasett. Arkivinnholdet i pakkene beskrives i arkivkatalogen.

For hver enkelt AIP, SIP eller andre datasett registreres informasjon som identifikatorer og navn, hvor de er plassert, informasjon om status, type, innhold, tilgjengelighet (innsynsløsning eller annet) og preserveringsoppgaver (issues). Registeret tilpasses til den enkelte bevaringsinsitusjons behov.

Et depotregister har flere bruksområder utover å gi oversikt over det som forvaltes i et digitalt depot. Det gir en mulighet til å arbeide systematisk over tid for å sikre at materialet i depotet forvaltes slik at det vil være langtidstilgjengelig. Det vil kunne gi verdifull informasjon om risiko som finnes i depotbestanden, og gi en mulighet til å kunne prioritere preserveringshandlinger ut fra dette. De identifiserte risikoelementene bør synliggjøres for hvert enkelt datasett i register over digitale ressurser, slik at de kan håndteres som del av den overordnede styringen og forvaltingen av depotfunksjonen.

Registeret vil kunne gjøre det mulig å gi en samlet fremstilling av forholdet mellom alle AIPer, SIPer og andre datasett og beskrivelse av disse i arkivkatalog, samt plassering for eventuelle innsynsløsninger. Dette er informasjon som for de fleste kan være vanskelig å få frem i arkivkatalogverktøyet ASTA.

En annen viktig funksjon for depotregisteret er å bidra til å synliggjøre det digitale depotet for resten av organisasjonen. Ikke bare hva det digitalt depotet inneholder, men også den arbeidsmengden og ressursene som det er behov for ved forvaltning av digitalt skapt arkiv. Det kan også benyttes i forbindelse med rapportering.

7.3 Krav til lagringsfunksjon

Det mest grunnleggende kravet til lagring av digitalt skapt arkiv er å forhindre at data går tapt, som et resultat av fysisk eller elektronisk skade på lagringsmedia, tyveri av data eller lagringsmedier, og utilsiktet eller ondsinnet sletting av data. Det primære virkemiddelet er å lagre flere identiske kopier av data. Det å lagre dem på ulike lokasjoner og begrense antallet personer som har tilgang til data, bidrar også til å styrke sikkerheten.

Videre må data overvåkes for å sikre at integriteten opprettholdes. Det vil si at data ikke endres gjennom utilsiktede eller ondsinnede handlinger, og at eventuelle endringer kan identifiseres. For å overvåke eventuelle endringer, er det vanlig å bruke sjekksum. Det er en sum som blir kalkulert av en algoritme basert på innholdet i én eller flere filer. Dersom en fil endres, vil sjekksum også endres. For å overvåke integriteten, må man gjennomføre regelmessige rekalkuleringer av sjekksum, og sammenligne den med den lagrede sjekksummen.

Sjekksum kan ikke brukes til å reparere innholdet som er skadet, men identifisere endring. Ved lagring av flere kopier av data er det mulig å gjenopprette integriteten i data når man identifiserer endringer i sjekksum.

Overvåking av integriteten til data bidrar også til å sannsynliggjøre at materialet i depotet er autentisk, det vil si at informasjonsinnholdet i materialet er uendret, og at det er hva det gir seg ut for å være. Over tid vil ikke overvåkning av integritet alene være tilstrekkelig for å ivareta autentisiteten. Da benyttes versjonskontroll for å bidra til å opprettholde tilliten til at materialet i depot er autentisk. Når filer konverteres til nye formater, som en del av forvaltningsoppgavene i et digitalt depot, skrives de om. I den prosessen endres filen, og informasjonsinnholdet blir i ulik grad påvirket av dette. For å sannsynliggjøre at den eventuelle påvirkningen av informasjonsinnholdet ikke er tilsiktet, bevares den opprinnelige versjonen i tillegg til den nye. I tillegg lagres informasjon som identifiserer endringen, gjennom identifikasjon av den som har utført endring og beskrivelse av endring. Integriteten til begge versjonene overvåkes videre.

7.4 Verktøy for overvåkning av lagret arkiv

Det er behov for verktøy for å overvåke lagret arkiv. Overvåkningen skal sannsynliggjøre at integritet og autentisitet er opprettholdt i det digitalt skapte arkivmaterialet som lagres og forvaltes i et digitalt depot. Verktøy som velges må i tillegg være tilpasset arbeidsmetodikk, ressurser og behov hos bevaringsinstitusjonen som skal bruke verktøyet. Det finnes flere alternative verktøy å velge mellom. Noen er laget spesifikt for digitalt skapt arkiv, mens andre er opprinnelig laget for andre formål, men har tilsvarende funksjonalitet. Her er noen eksempler:

Bagger[7] er et av flere verktøy som er laget etter spesifikasjon av BagIt-formatet.[8] BagIt-formatet er utviklet for å overvåke at det ikke skjer endringer i digitalt materiale ved transport/flytting eller lagring. Det benyttes sjekksum for å overvåke integriteten til et materiale gjennom opprettelsen og kontroll av en bag. En bag består av mapper og filer som skal bevares, samt tekstfiler som legges til av Bagger. Tekstilene inneholder bl.a. oversikt over alt innhold i baggen, sjekksum for alle enkeltfiler og hele baggen samlet, og kontekstinformasjon som legges til av Bagger eller bruker.

Det må regelmessig gjennomføres ny kalkulering av sjekksum for hver enkelt bag for å kontrollere at integriteten er ivaretatt. BagIt-formatet har ingen støtte for versjonering av filer i en bag, men det finnes metoder for å kunne inkludere dette.[9]

Som et verktøy for lagring i et digitalt depot må Bagger (BagIt) anses å være et enkelt verktøy egnet for depot med begrenset mengde data, og kanskje primært som del av tiltak i forbindelse med overgang eller planlagt innføring av en annen løsning.

ESSArch[10] er brukt av flere bevaringsinstitusjoner i Norge, og er tilpasset spesifikasjoner fra DIAS-prosjektet.[11] ESSArch er også et versjonskontrollsystem, men baserer seg på å pakke arkivdata inn til en eller flere pakkefiler (tar-fil), omtalt som en arkivpakke, tilsvarende en bag i Bagger eller et repo i Subversion. Arkivpakkene må hentes og pakkes ut før man når innholdet. Endringer som gjennomføres, f.eks. konvertering av filer, vil gi en ny versjon av hele arkivpakken når den returneres til ESSArch. Alle endringer utført av depottjenesten loggføres. ESSArch har støtte for lagring og versjonskontroll på flere separate lokasjoner, og har støtte for lagring til både disk og tape. Det er også støtte for arkivspesifikke standarder som METS[12] og Premis.[13] Det er noen tilleggsverktøy til ESSArch som kan benyttes for å sende og motta arkivpakker, ikke ulik funksjonaliteten i Bagger.

Subversion[14] er et versjonskontrollsystem (tilsvarende GIT, Perforce m.fl.). Det er åpen kildekode som del av Apache Software Foundation. I Subversion blir en datamengde organisert i et «*repo*» (*repository*). Et repo

[7] Bagger: https://github.com/LibraryOfCongress/bagger
[8] BagIt: http://www.digitalpreservation.gov/series/challenge/data-transfer-tools.html
[9] https://stacks.wellcomecollection.org/how-we-store-multiple-versions-of-bagit-bags-e68499815184
[10] EssArch: https://www.essarch.org/
[11] DIAS: https://www.arkivverket.no/forvaltning-og-utvikling/regelverk-og-standarder/dias-prosjektet-digital-arkivpakkestruktur
[12] METS: http://www.loc.gov/standards/mets/
[13] PREMIS: https://www.loc.gov/standards/premis/
[14] Subversion: https://subversion.apache.org/

består av en mappe som inneholder data som skal bevares og overvåkes. Subversion genererer sjekksum for alle enkeltfiler for å overvåke integritet i materialet, og holder oversikt over versjonshistorikken til filer for å sikre autentisitet. Alle endringer depottjenesten utfører på materialet loggføres som del av grunnfunksjonaliteten til Subversion. Subversion og andre tilsvarende versjonskontrollsystem er utviklet for å holde oversikt over versjoner av kode i forbindelse med programvareutvikling. De fleste har mye funksjonalitet utover det som er relevant i et arkivdepot.

Under vil vi vise de få og enkle funksjonen som benyttes for å ivareta digitalt skapt arkiv lagret i et digitalt depot hvor Subversion blir brukt.

7.5 Versjonskontrollsystem som digitalt sikringsmagasin

I Bergen byarkiv tester vi ut Subversion som digitalt sikringsmagasin. Det er en enkel og fleksibel løsning. Et versjonskontrollsystem tillater at man oppdaterer enkeltfiler og har flere versjoner av samme filen. I motsetning til systemer som Essarch og Archivematica, så kan man oppdatere enkeltfiler uten å opprette en helt ny arkivpakke. Det er følgelig mye enklere å oppdatere en pakke, og man fordobler ikke langringsplassen når man gjør en liten endring på en fil.

Subversion har også funksjonalitet for å sjekke integriteten til alle filer - noe man da f.eks. kan gjøre hver natt. Det bør også tas backup av repoene, slik at man er sikret hvis noe skulle skje.

Man kan sette opp brukerstyring på Subversion-serveren, slik at kun autoriserte personer får tilgang til et spesifikt repo. Det er også mulighet med detaljert brukerstyring på mappenivå innen et repo.

Det finnes også andre versjonskontrollsystemer som kan være aktuelle å bruke, som Git med LFS (støtte for store binære filer), eller Perforce, som har overtatt for Subversion i mye av spillutvikler-miljøet. Grunnen til at vi valgte Subversion, var for det første at det er åpen kildekode, samt at det har bedre støtte for store binære filer enn Git. Git har fått støtte for slike med utvidelsen LFS, men det krever at man aktivt markerer alle filer som skal håndteres av LFS. Dessuten har Git mulighet for omskrivning av historikken i et repo, og det er ikke veldig heldig i et digitalt sikringsmagasin.

7.5.1 Kvalitetssikring av innhold og innlegging i digitalt depot

Når man klargjør materialet for å legges inn i depot, kan det være hensiktsmessig å følge en standard for hvordan pakken struktureres, og hvilke filer som skal være med. Fordelen med å ha en standard er at det blir lettere å finne igjen informasjon, og langt lettere å automatisere slik uthenting og oppdatering.

Hvis man bruker et versjonskontrollsystem, kan man lett implementere såkalte “pre commit hooks”, dvs. at det kjøres skript før filer og endringer legges inn. Disse skriptene kan da sjekke om strukturen er riktig, og om alle filer som skal være der, er der, samt om disse filene inneholder den informasjonen de skal. Dersom skriptet finner at noe er feil, mislykkes innleggingen. Dermed er man sikret at et uttrekk som er lagt i digitalt sikringsmagasin, inneholder alt det skal, og er strukturert på riktig måte.

Når man har ferdigstilt en arkivpakke med alle nødvendige filer, kan man opprette et repo, og sjekke inn alle filene. Arkivpakken kan da betraktes som ferdig avlevert. I et versjonskontrollsystem er arkivpakken altså bare en samling med filer som til sammen utgjør et «repo».

7.5.2 Preserveringsplanlegging

For å opprettholde lesbarhet av dokumenter over tid, er det nødvendig å sørge for at filer med filformater som fases ut, blir konvertert til nytt format.

Som nevnt ovenfor (Jf. kapittel “Uttrekk fra opprinnelig system”) produserer vi en tsv-fil med informasjon om hver enkelt fil, inkludert filformat til konvertert fil. Hvis denne tsv-filen har samme navn i alle repoene

og ligger på en standard plassering, kan man enkelt med et skript hente denne ut fra hvert repo og søke gjennom den for å finne alle filformater som ligger der. Å hente ut enkeltfiler er enkelt i Subversion.

I Bergen byarkiv benytter vi en enda enklere løsning for å holde oversikt over filer og format. Vi kjører et lite skript som gjennomgår alle filene i repoet, og oppretter en fil "innhold.txt" med sti til hver fil på egen linje. Dermed får vi også med dokumentasjonsfiler. En slik fil er veldig enkel å gå gjennom med et skript for å finne filer som må konverteres på nytt. Da brukes bare filutvidelsen til å utlede filtypen. Genereringen av filen kan legges til en pre-commit-hook, slik at man ikke behøver å huske på å generere den på nytt når man er inne og gjør noen endringer.

Filene med sql for å gjenopprette opprinnelig database samt opprette innsynsbase skal ikke behøves konverteres, eller det kan gå veldig lang tid før man behøver å tenke på det. Når man bruker ISO SQL (jf. kapittel om uttrekk) vil man kunne gjenskape databasene i veldig lang tid framover.

7.5.3 Preserveringshandlinger

Når man skal konvertere filer på nytt i et versjonskontroll-repo, henter man bare ut de filene med det formatet man er ute etter. Dette kan enkelt gjøres med et skript som går gjennom filoversikten ("innhold.txt"), henter ut en og en fil ut og konverterer dem. Til slutt legges filene tilbake til repoet, og det logges hva som er gjort i en "commit message". Filoversikten i "innhold.txt" oppdateres automatisk av en pre-commit-hook.

Når Bergen byarkiv bruker et versjonskontrollsystem, kan vi altså når som helst oppdatere noen filer, og lagre dem tilbake til repoet. Dette tar bare den ekstra lagringsplassen som de konverterte filene tar.

Kapittel 8

Verktøy

8.1 PWCode og PWLinux

PWCode er en programvare utviklet ved Bergen byarkiv for å ivareta våre behov for å ta uttrekk og normalisere data for lagring i digitalt depot.

Programvaren kan hente ut de fleste data fra datasystemer, og lagre disse som normaliserte data i en arkivpakke. Per i dag kan en eksportere alle data fra et system som lagrer dataene i en relasjonsdatabase eller som filer på disk. Støtte for å hente ut data fra NOSQL-databaser og skytjenester er planlagt.

PWCode er en del av prosjektet Preservation Workbench, som samler verktøy brukt til håndtering av digitalt skapt arkivmateriale.

Den andre hoveddelen av Preservation Workbench er PWLinux, som er et automatisert oppsett av Linux, med alt en trenger installert for normalisering og testing av data med PWCode.

Programvaren tilhørende Preservation Workbench er tilgjengelig for nedlasting her: https://github.com/Preservation-Workbench

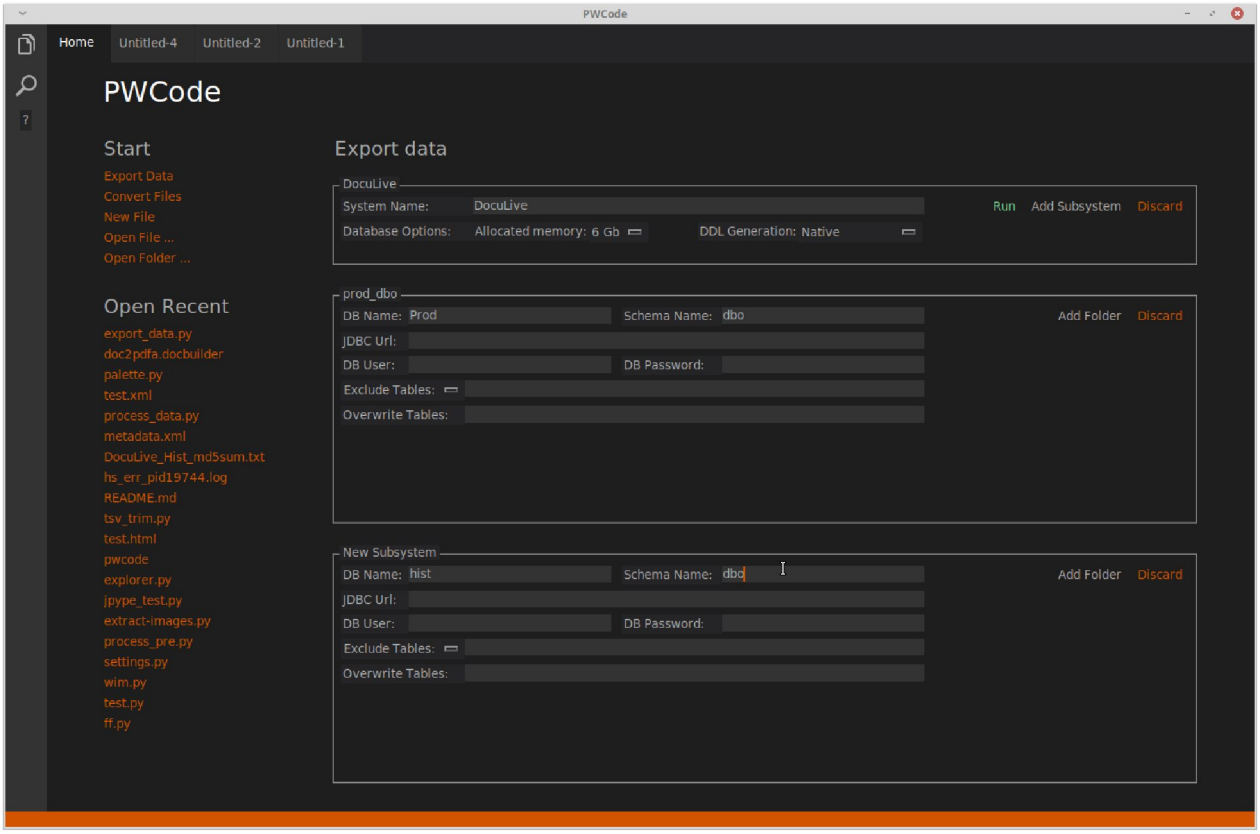
8.1.1 Eksport av data

PWCode kan brukes til å eksportere data fra både Windows- og Linux-maskiner, uten å måtte installeres, og uten å kreve at databasedrivere eller annet er installert på maskinen. Det holder å laste ned PWCode som zip-fil fra https://github.com/Preservation-Workbench/PWCode/archive/master.zip, åpne zip-fil til mappe, og starte programmet.

Første gang en starter programmet vil nødvendig python- og java-kode og databasedrivere lastes ned. Etter at dette er gjort, kan en kopiere mappen hvor som helst, og kjøre programmet direkte fra mappen uten installasjon. Det virker også når mappen er på en ekstern disk eller på filserver.

Når en starter et datauttrekk med PWCode, blir det først opprettet en mappe under 'Projects' med samme navn som systemet en henter data fra. I denne mappen blir det så automatisk opprettet standard mappestruktur for en arkivpakke. Det ferdige datauttrekket blir dermed til slutt en nesten helt ferdig arkivpakke som kan legges i digitalt depot.

Data fra relasjonsdatabaser blir i første omgang kopiert til lokale H2-databaser i arkivpakken (hvert databaseskjema blir én fil). På denne måten gjør en det mulig å mye enklere behandle data videre hos depot i etterkant enn hvis data ble eksportert til filer på disk med én gang. I tillegg brukes det mindre minne enn ved eksport til disk, og uttrekk som ellers ikke hadde vært mulig fra maskin med lite tilgjengelig minne, kan derfor likevel gjennomføres. Hvis prosessen likevel stopper opp pga minnemangel, som kan skje med tabeller



Figur 8.1: PWCode

med store mengder binære data, kan en starte opp uttrekk på nytt, og PWCode vil fortsette eksporten fra der det stoppet sist.

Denne metodikken har i tillegg den fordelen at data i mye mindre grad endres i første eksport enn ved eksport til data på disk (H2 er en ren java database og data blir dermed ikke transformert mer enn det uansett ville blitt når en bruker JDBC-databasedriver for å aksessere dataene). Dette er en stor fordel med datatyper som blob, clob, LONG RAW m.m., hvor det spesielt for en del eldre databaser kan være behov for tilpasning av eksport til disk, for at data skal beholdes på et riktig og lestbart format.

Vi får også en mye mer smidig prosess, hvor en har mulighet til å prøve og feile på eget kontor, heller enn å ta opp tid hos arkivskaper. Selve uttrekket av data hos arkivskaper tar dermed veldig lite tid.

I tillegg til å hente ut alle data fra en database, blir det også automatisk hentet ut en beskrivelse av opprinnelig databasestruktur som metadata.xml. Denne kopieres til arkivpakke sammen med H2-databasen. Denne filen inneholder all informasjon en finner i metadata.xml i en SIARD-pakke, pluss en god del mer.

Data i form av filer på disk blir eksportert i form av pakkede tar- eller wim-arkiver (én pr. kopiert mappestruktur). Tar brukes når data hentes fra en Linux-maskin. Wim-formatet brukes når data hentes fra en Windows-maskin, fordi kun dette arkivformatet støtter alle fil-metadata på NTFS- filsystemet (og nyere varianter) som brukes av Windows. Formatet kan i tillegg åpnes som et disk image i etterkant (i motsetning til å pakkes ut) på både Windows og Linux. En beholder dermed alle opprinnelige metadata på filene en har eksportert, og en kan gjøre den tidkrevende (men automatiske) jobben med å hente ut alle metadata i depot heller enn hos arkivskaper, uten å oppleve at metadata har blitt vekke eller har blitt endret på veien (som kunne vært tilfelle med annen metodikk). I tillegg har denne metodikken den fordelen at en ikke får problemer med begrensninger på Windows i form av lange filbaner (maks 255 tegn).

Når alle data er ferdig eksportert, blir arkivpakken (mappestrukturen med alle eksporterte data) pakket som et tar-arkiv, og det genereres en sjekksum for tar-filen.

8.1.2 Arkivbegrensning

PWCode har begrenset funksjonalitet for arkivbegrensning, da dette delvis er en manuell operasjon. Tabelldata kan arkivbegrenses per tabell manuelt ved å slette en eksportert TSV-fil i forkant av normalisering av data (se beskrivelse av normalisering under). Ved normalisering av data detekteres dette, og det dokumenteres i metadata.xml at en arkivbegrensning er blir gjort. En kan i etterkant - hvis ønskelig - oppdatere standard tekst for dette i metadata.xml med en mer presis begrunnelse.

Mer støtte for arkivbegrensing i PWCode er planlagt. Det vil kunne innebære at man i kartleggingen og dokumentsjonen av databasen markerer at tabeller skal kasseres, og så kan PWCode fange opp dette og holde tabellen utenfor ved generering av arkivpakke.

8.1.3 Normalisering av data

Med normalisering av data menes at data konverteres til formater som er mest mulig programuavhengige og anvendbare for framtiden. For normalisering av data kreves en full installasjon av Preservation Workbench (PWCode installert på PWLinux), da bare denne har installert de nødvendige programmene for konvertering av filer m.m. Støtte for Windows 10 (via WSL2) er planlagt.

Når eksporterte data er overført fra arkivskaper til depot, åpnes tar-arkivpakken med PWCode, og det generes en ny sjekksum av data som sjekkes mot sjekksum generert hos arkivskaper, for å verifisere at data ikke har blitt endret på veien. Det vil f.eks. kunne skje ved kopiering av data mellom forskjellige operativsystemer, filsystemer, eller ved overføring av data med FTP uten bruk av 'binary transfer mode'.

I etterkant av verifisering av data med sjekksum, lager PWCode en sikkerhetskopi av eksporterte rådata (SIP-en). Så starter den automatiske normaliseringsprosessen, hvor det lages ferdig normalisert AIP fra original SIP. En kan gå tilbake til rådata og starte prosessering på nytt ved behov.

Formatene vi har valgt å bruke for normalisering av tabell-data fra relasjonsdatabaser, er TSV og SQL (ISO/IEC 9075). Selve dataene lages i TSV-filer (én pr tabell) og SQL DDL (Data Definition Language) brukes til å dokumentere datastrukturen.

SQL (Structured Query Language) er den dominante metoden for manipulering av data i en relasjonsdatabase, men det vil typisk være forskjeller på hvordan dette er implementert i forskjellige databasemotorer. ISO SQL er en databasemotor-uavhengig standardisering av SQL.

SQL DDL blir automatisk generert av PWCode fra `metadata.xml`-filen som ble generert i selve uttrekket hos arkivskaper. I tillegg til å dokumentere datastrukturen, har DDL-koden den fordelen at det også er fungerende SQL-kode, som kan brukes til å direkte gjenskape opprinnelig database i en hvilken som helst databasemotor (relasjonsdatabase) som har tilstrekkelig god støtte for SQL-standarden med en vanlig databaseklient.

De fleste relasjonsdatabaser i utstrakt bruk har per i dag ganske god støtte for den begrensede delmengden av SQL-standarden som vi trenger for våre behov. De beste i så måte er PostgreSQL og SQLite, som støtter alt vi trenger. For MySQL, Oracle og MSSQL er støtten god nok til at det kun trengs en eller to 'søk/erstatt' mot den genererte sql-filen for å få kode som kan brukes direkte. For MSSQL er dette f.eks. at datatypen TIMESTAMP byttes ut med DATETIME2. Det er altså kun for på datatyper det finnes gjenværende avvik fra SQL-standarden. PWCode håndterer disse avvikene ved å genererer automatisk varianter for de databasene dette gjelder, i tillegg til den rene ISO SQL-varianten. For framtiden er det god sjanse for en ikke trenger dette mer da utviklingen de siste årene er at stadig større deler av SQL-standarden implementeres i de databasemotorene som har hatt en del mangler på dette. For få år siden ville det ikke vært mulig å kunne støtte MySQL og MSSQL på denne måten i det hele tatt pga veldig store mangler i disse databasene.

Det aller vanligste formatet pr. i dag for lagring av tabelldata, er CSV-formatet. Dette er også blant de formatene Arkivverket godkjenner. Spesifikt godkjennes CSV iht. RFC 4180. Som standard har RFC 4180 mange problemer, men de tre største er at den ikke tillater unicode (som vi må ha), at den ikke tillater Unix linjeskift (LF), samt at en ikke kan garantere seg mot feil ved innlesing av filer maskinelt, siden separasjonstegnet (typisk komma eller semikolon) også er tillatt som data i kolonner. Hvis alle implementeringer av CSV i kode tolket dette helt likt, ville dette vært et overkommelig problem, men tester med et stort utvalg databaseklienter viser at dette ikke er tilfelle.

Vi trengte derfor et alternativ som garantert kunne leses maskinelt (enklest garantert ved at separasjonstegnet aldri kan være del av innhold i kolonne), og som ellers har de samme egenskapene som CSV. Heldigvis finnes det allerede i form av TSV-filer med unicode encoding (UTF-8 i vårt tilfelle). TSV-filer bruker tab-karakteren som separasjonstegn, og det er ikke tillatt å ha tab-tegn som del av innhold i kolonne. PWCode bytter derfor ut eventuelle tab-tegn i selve dataene med mellomrom ('space') ved eksport av data fra H2-database til tsv-fil på disk. Fordi TSV som format er så enkelt, er det som godt som aldri forskjell på hvordan dette er implementert i kode, og vi får derfor ikke problemer ved opplasting av data til en database med en vilkårlig databaseklient.

Hvis en velger å lagre tabelldata i XML (brukt bl.a. i Siard-pakker og Noark-uttrekk), får en umiddelbart to problemstillinger: Lagrede data tar opp mer plass på disk (enn f.eks. CSV), og en er avhengig av spesialverktøy/spesialkode for å kunne laste opp data til en ny database, som en må, både for testing av data, og for å kunne lage en innsynsløsning. En kunne også valgt JSON, YAML e.l. men de ville alle hatt de samme to problemstillingene.

Binære data (blob-er) og store tekstfelt (clob-er) blir i PWCode som hovedregel eksportert til separate filer på disk, med automatisk genererte unike navn, som angir kolonne og tabell de ble eksportert fra. Det er planlagt mer finkornet sjekk av clob-felt, for i noen tilfeller (når tekstmengden er innenfor det som kan oppbevares i et standard SQL felt for tekst-data i alle støttede databasemotorer) å beholde disse som tekstdata i TSV-filen. Denne eksporten skjer før normalisering av filer på disk, slik at både filer opprinnelig på disk, samt filer eksportert fra felt i database, blir normalisert.

I forkant av selve normalisering av filene blir det først gjort en automatisk virussjekk. Hvis data har blitt produsert helt fram til uttrekks-tidspunkt, bør en legge inn en karantenetid for rådata og utsette normaliseringprossen. I neste omgang blir alle filene sjekket med programvare som identifiserer filtypen. Den gjør en grundig sjekk av hver fil, for nøyaktig å kunne sjekke hvilket filformat en fil er. Man kan nemlig ikke stole på

filendelse alene. I tillegg eksporterer PWCode alle metadata på filene til en TSV-fil som legges i arkivpakken. Alle metadata på filer er dermed sikkert bevart slik de var opprinnelig. Slike metadata er umulig å garantere mot endringer ved vanlige filoperasjoner som kopiering m.m.

I neste omgang blir alle filer som trenger det, basert på informasjon i TSV-fil, konvertert automatisk til arkivformat. TSV-filen blir oppdatert automatisk av PWCode med informasjon om konvertering av filer til arkivformat, og hva navn og filbane på arkiv-varianten av filen er. Denne TSV-filen kan også danne utgangspunkt for en innsynsløsning for uttrekk av kun filer i mappestruktur (som ofte er tilfelle for privatarkiv) på samme måte som generert SQL DDL danner utgangspunkt for en innsynsløsning for data som opprinnelig var i en database.

For valg av arkivformat følger vi som hovedregel anbefalinger fra Arkivverket for arkivformater, men i de tilfeller disse er mangelfulle ser vi i stor grad til 'Library of Congress Recommended Formats Statement'. I tillegg vil vi i noen tilfeller velge et arkivformat over et annet på basis av hva som fungerer best i en innsynsløsning. Dette kan også enkelt endres ved behov i PWCode.

8.1.4 Verifisering av normaliserte data

For testing av normaliserte data kreves også en full installasjon av Preservation Workbench.

Denne testen er hovedsaklig for tabelldata, og for å teste at tidligere generert SQL DDL virker på alle støttede databasemotorer. Per i dag kjøres den automatisk for hvert uttrekk. På et senere tidspunkt, når nok systemer har blitt testet og vi føler oss sikre på at alle eventualiteter er håndtert riktig av koden, vil vi vurdere å kutte dette steget.

Tester kjøres automatisk mot følgende databasemotorer: Postgresql, MySQL, Oracle, MSSQL og SQLite. Dette er per i dag de mest brukte databasemotorene i verden, og dermed de viktigste å teste mot. I alle tilfeller genereres ny database, og data importeres med standardverktøy for databasemotoren. Dette for å sikre at data og databeskrivelse i tillegg til å være systemuavhengige også er verktøyuavhengige.

8.1.5 Ferdigstilling av arkivpakke

Etter at validering av data er gjort, sitter en igjen med en ferdig arkivpakke som kun mangler innlegging av arkiv- og aktørbeskrivelse, innhentet systemdokumentasjon, og dokumentasjon av innsynsdatabase. Når dette er gjort manuelt kan pakken legges inn i digitalt depot.

8.1.6 Planlagt støtte for andre formater

Ved tidligere bruk av SIARD hadde vi store utfordringer med spesielt opplasting av store uttrekk, og opplasting til annen databasetype enn den uttrekket var gjort fra opprinnelig. Men siden SIARD er såpass utbredt i arkiv-miljøet, vurderer vi å legge inn støtte også for SIARD-uttrekk i PWCode. Foreløpig plan er å generere SIARD-uttrekk som et siste trinn i dagens metodikk, og i etterkant av at data allerede er ferdig normalisert med TSV og SQL. En vil dermed ha luket ut disse problemene i forkant, og kan være sikker på at SIARD-pakken lar seg laste opp til en ny database.

Vanlig SIARD-metodikk er at programvaren endrer på data (tilpasset databasemotor en laster opp til) først i det en laster opp. Og det er på dette punktet en vil kunne oppleve problemene nevnt over. Vi endrer dermed bare to ting: Vi normaliserer (og tester) data i forkant, og vi normaliserer bare til minste nødvendige delmengde av SQL (SIARD støtter hele SQL:2008). Men siden vi bruker en delmengde får en fortsatt en pakke som støtter standarden. Vi har dermed forenklet oppgaven pakken må løse, ved å kutte ned på antall variabler som må tas høyde for. Oppgaven til SIARD-programvaren en bruker for å laste opp data til en ny database blir dermed langt enklere, og sjansen for feil mye mindre.

For enda bedre standard-støtte er det planlagt å legge inn støtte for Data Package-formatet, som utenfor arkiv-miljøet er langt mer utbredt enn SIARD (https://frictionlessdata.io/data-package/#tabular-data-package).

Det eneste som mangler i PWCode for å støtte denne standarden, er at det for hvert uttrekk også genereres en `datapackage.json`-fil. Denne filen innholder samme informasjon som vi allerede har lagret i arkivpakken i form av `metadata.xml`. En trenger dermed bare legge inn kode i PWCode for å konvertere metadata.xml til en json-fil som er bygget opp slik det kreves av datapackage standarden. En oppnår dermed at våre uttrekk fra databaser er lagret på en måte som støtter flere standarder samtidig, uten å ha duplisering av lagrede data. Det er på samme måte som for SIARD støtte for automatisk opplasting til databaser med standard programvare for datapackage. En har dermed flere alternativer for hvordan en går fram for å laste opp data til en ny database for å lage DIP eller annet.

8.2 Universal Relational Database (URD)

URD er en programvare for å kunne vise fram og registrere data fra enhver relasjonsdatabase. Den ble startet utviklet ved Oslo Byarkiv, og er nå tatt i bruk ved Bergen byarkiv for å dokumentere alle uttrekkene våre. I tillegg til framvisning av data, kan den også brukes til å analysere hvordan databaser henger sammen, og hvilke deler som har vært i bruk.

Kildekoden finnes her: https://github.com/fkirkholt/urd

8.2.1 Analyse av database

Før man kartlegger opprinnelig database er det en fordel at mest mulig "støy" er fjernet. En database er ofte veldig stor, med hundrevis av tabeller, og hvor det også kan være flere titalls kolonner i mange av tabellene. Samtidig er det ofte mange tabeller og kolonner som er helt tomme. Da er det en fordel om slike tomme tabeller og kolonner fjernes fra databasen før man kartlegger den.

Hvis URD finner at en tabell har ingen eller bare én post, genereres det en "drop table"-setning, slik at man kan velge å fjerne disse tabellene. Samtidig genereres det "drop column"-setninger, slik at eventuelle kolonner som refererer til disse tabellene også fjernes.

URD kan også finne ut hvilke kolonner som har vært i bruk, og hvilke som bare har vært i sporadisk bruk. Man setter en terskelverdi for hvor stor andel av postene som skal ha verdi i en kolonne, og så skjuler URD de kolonnene som kommer under denne terskelverdien. Slik kan man raskt få oversikt over den mest brukte informasjonen i systemet. Ved å justere terskelverdien, kan man bestemme hvor mye en kolonne skal ha vært i bruk for at den vises, og slik kan man velge hvor stor del av informasjonen man forholder seg til. Det kan være lurt å begynne og sette en høy verdi, og så sette den stadig lavere, for å se på den mindre brukte informasjonen etter hvert. For kolonner som har vært lite i bruk, må man vurdere hvor relevante disse er å ta med i en innsynsløsning.

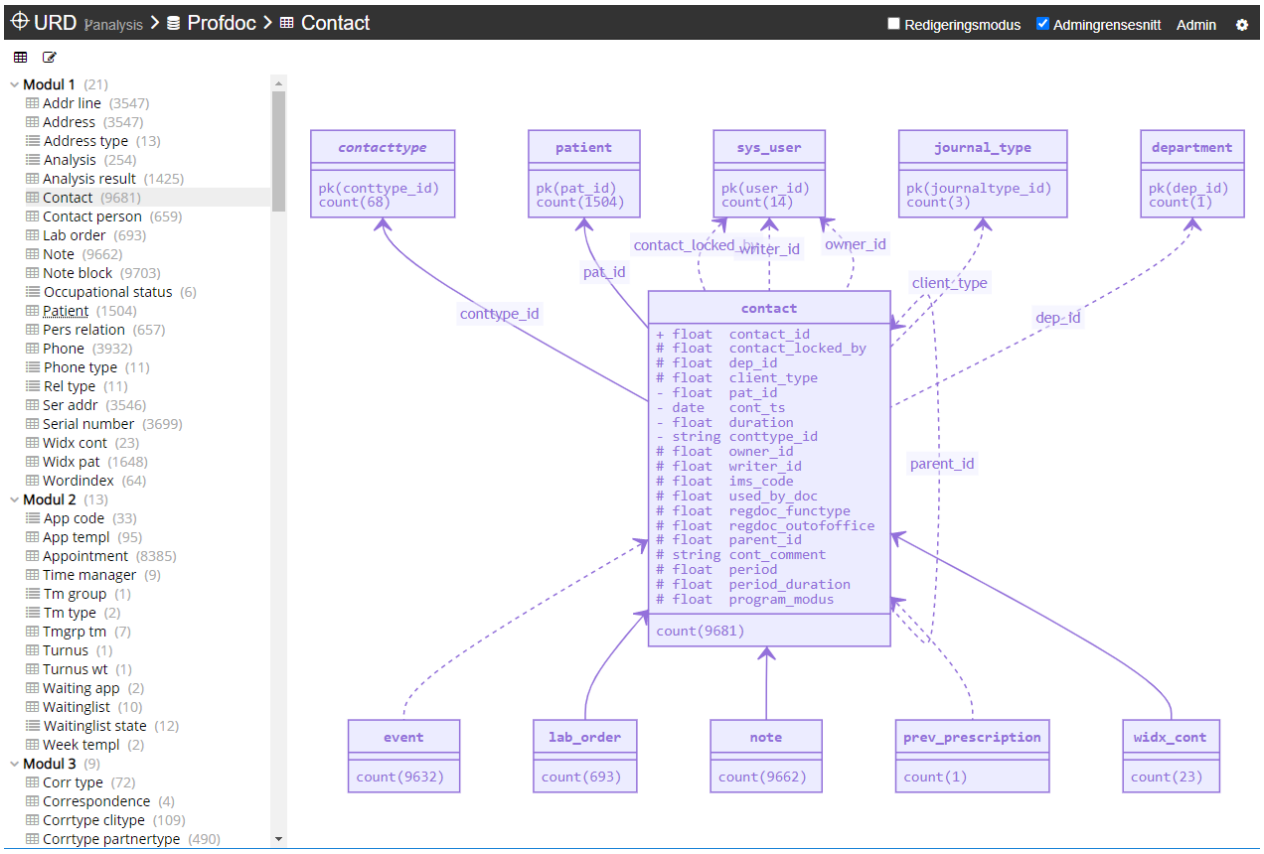
En database kan også ha mange referansetabeller (også kalt oppslagstabeller), hvor mange av postene ikke har vært i bruk. URD analyserer koplingene til disse referansetabellene, og genererer "delete"-setninger for å fjerne de verdiene som ikke har vært i bruk. Slik kan man rydde opp i databasen, og fjerne ubrukte poster i referansetabeller. Da ser man også hvilke verdier som faktisk har vært i bruk.

URD kan brukes til å kartlegge strukturen i en database for å se hvordan tabeller henger sammen. Dette er veldig nyttig, da det er vanskelig å få oversikt over store databaser.

Ved å se på hvilke fremmednøkler som finnes i databasen, kan URD finne hvilke tabeller som hører sammen, og gruppere disse i moduler. Man kan også justere hvilke tabeller man ønsker skal være med og definere moduler, og dermed styre hvilke moduler man får. Det finnes jo tabeller som brukes i hele systemet, uten at de egentlig inngår i spesielle moduler.

Ved å dele inn i slike moduler, får man lett synliggjort hvilke som er de viktigste tabellene, dvs. hvilke tabeller man ønsker å ta med i en innsynsløsning.

URD viser fram relasjonene mellom tabeller grafisk, dvs. i et slags ER-diagram. Man kan velge å vise fram kun tabellene innenfor en bestemt modul.



Figur 8.2: ER-diagram i URD

Mange databaser mangler fremmednøkler, slik at man ikke ser hvilke relasjoner som finnes mellom dataene. Da vil ikke URD greie å finne ut hvordan basen henger sammen. Da måtte man eventuelt gjennomgå databasen og finne mulige kandidatnøkler basert på hvilke verdier som er brukt i kolonnene. Ellers kan også views og triggere inneholde opplysninger om hvordan data er bundet sammen. Disse kan analyseres manuelt, eller man kan forsøke å finne disse relasjonene automatisk. Men en slik funksjonalitet er ikke bygd inn i URD ennå.

8.2.2 Innsynsløsning og dokumentasjon

URD kan brukes til å vise fram data fra enhver relasjonsdatabase. Man må da generere en json-fil, som brukes til å bestemme hvordan basen skal vises fram. Denne filen kan URD generere, og så kan man redigere den for å bestemme hva som skal vises fram, og hvordan basen skal presenteres.

Når man skal lage en innsynsløsning, er det greit å bruke det vi har valgt å kalle "selvdokumenterende databaser". Det er relasjonsdatabaser som er laget etter noen bestemte regler, som gjør at URD kan lage en innsynsløsning av basen kun basert på selve databasestrukturen. En slik base vil fungere veldig godt som dokumentasjon av opprinnelig database, da den tar med kun vesentlig data som er brukt i opprinnelig database.

I reglene vi har satt opp for hvordan selvdokumenterende baser skal fungere, har vi forsøkt å bygge på ofte brukte designprinsipper for databaser. Reglene skal sørge for god databasedesign, samtidig som de brukes for å dokumentere databasen.

Tabellnavn og kolonnenavn bestemmer ledetekst til tabeller og kolonner. Disse kan grupperes ved at man bruker felles prefiks. F.eks. vil kolonnene `periode_fra` og `periode_til` grupperes under overskrift "Periode", og ledetekst for feltene blir "Fra" og "Til". Tabeller grupperes på samme måte i moduler under felles overskrift.

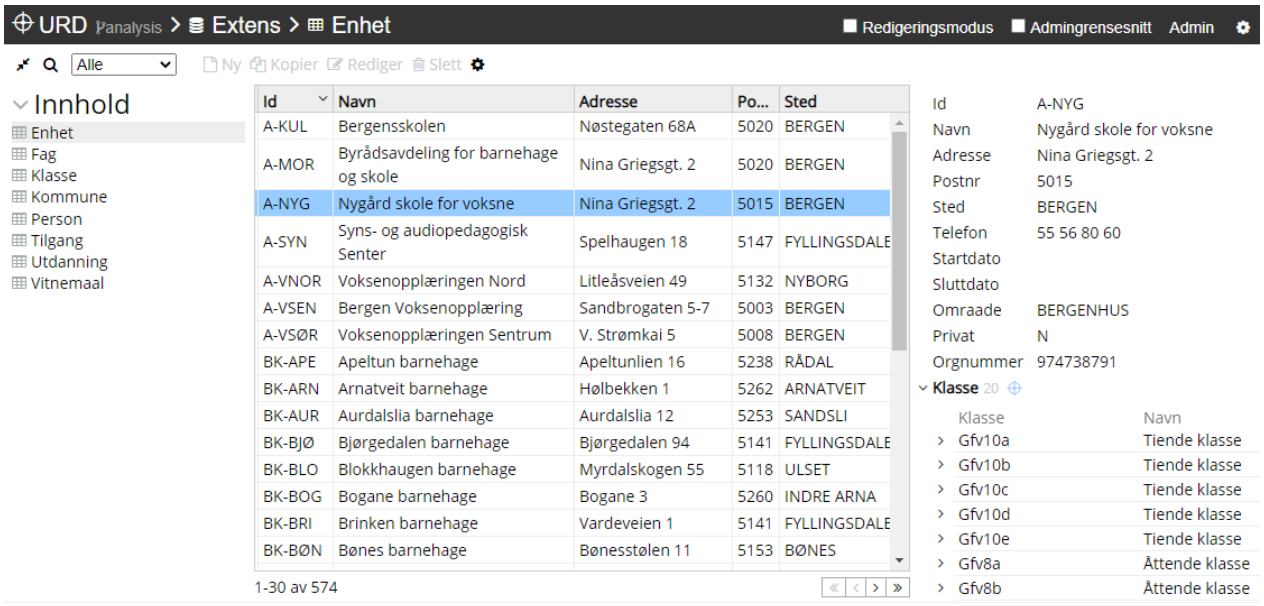
Prefikser kan forkortes. I så fall må man ha en tabell `meta_terminology` som beskriver hva prefikset betegner. Denne tabellen er nyttig uansett for å forklare begreper brukt i et fagsystem. Her kan man også legge inn beskrivelser som dukker opp som verktøytips når man holder musepekeren over en ledetekst.

Når man skal vise fram en bestemt tabell i URD, må man velge ut hvilke kolonner som skal vises. Dette bestemmes av en indeks med navn på følgende format: `<tabellnavn>_grid_idx`. Indeksen fører også til at søk går vesentlig raskere, for spørringen går da kun mot indeksen.

Relasjoner finnes vha. fremmednøkler, og her brukes også indeks-navn til å bestemme ledetekst for relasjonene.

URD skal også tilrettelegges for å håndtere tilgangsstyring. Dette kan gjøres ved å ha en index `<tabellnavn>_access_idx` på en kolonne. Da vet URD at denne kolonnen brukes til å styre tilgang. Kolonnen kan f.eks. angi brukergruppe. Da finner URD tabellen over brukergrupper, og deretter lokaliseres tabellen som knytter brukere og brukergrupper sammen, ved at man i denne tabellen har en index `<tabellnavn>_user_idx`. Dermed ved URD hvordan spørringen mot en tabell skal være for å filtrere bort de postene som brukeren ikke har tilgang til. Dette åpner for en veldig fleksibel måte å styre tilganger på. Man kan gjenbruke tilgangstabell i opprinnelig database, og man kan styre ulike tilganger for ulike tabeller. Denne funksjonaliteten kommer i en senere versjon av URD.

Se ellers dokumentet `innsynsdatabaser.md` i Github-repoet til URD (github.com/fkirkholt/urd) for hvordan selvdokumenterende databaser skal struktureres.



Figur 8.3: Innsyn i URD