# Vedlegg 3A:

# Bevaring av fagsystemer

# Innhold

**1 INNLEDNING ..... 3**

1.1 HVA ER ET FAGSYSTEM ..... 3
1.2 FAGSYSTEMER MED DATA ORGANISERT I FLATE SEKVENSIELLE TEKSTFILER ..... 3

**2 BEARBEIDING AV FLATE SEKVENSIELLE TEKSTFILER ..... 3**

2.1 ENDRINGER I POSTLAYOUT OVER TID ..... 4
2.2 FJERNING AV SLETTEDE POSTER ..... 4
2.3 BYTTE TEGNSETT FRA EBCDIC TIL ASCII ..... 4
2.4 SKIFTE FRA 7-BITS TIL 8-BITS TEGNREPRESENTASJON ..... 4

**3 DOKUMENTASJON AV DATAFILER ..... 5**

3.1 TEKNISKE METADATA ..... 5
3.2 ANNEN SYSTEMDOKUMENTASJON ..... 5

**4 HVA KREVES FOR Å KUNNE BEVARE ET FAGSYSTEM? ..... 6**

4.1 TEKNOLOGISKE BETINGELSER ..... 6
4.2 KOMPETANSE ..... 6

# 1 Innledning

## 1.1 Hva er et fagsystem

Mange private arkivskapere lagrer informasjon i databaser som kan være alt fra enkle registre til fagsystemer med avansert forretningslogikk. Et fagsystem kan være et spesialsystem som er laget for å løse en bestemt arbeidsoperasjon hos én arkivskaper, eller et standardsystem som brukes av mange arkivskapere, som for eksempel regnskaps- eller lønnssystemer.

Mange depotinstitusjoner har opparbeidet kompetanse om avleveringer av fagsystemer fra offentlig sektor, særlig på store og kompliserte systemer. Arbeidet med slike avleveringer fra privat sektor må støtte seg til denne kompetansen. Prosjektgruppen har derfor valgt å se bort fra bevaring av store og kompliserte fagsystemer innenfor rammen av prosjektet.

## 1.2 Fagsystemer med data organisert i flate sekvensielle tekstfiler

Fokus i prosjektet har vært fagsystemer der dataene er organisert i flate sekvensielle tekstfiler. Denne metoden for organisering av data var utbredt før det ble tatt i bruk relasjonsdatabaser, og mange eldre fagsystemer inneholder data lagret på denne formen.

Lagring av flate sekvensielle tekstfiler er en enkel måte å organisere data på. Her samles data som hører sammen i registre, for eksempel kunderegister med alle opplysninger om kunder, eller vareregister med alle opplysninger knyttet til varer. Mange av fagsystemene med flate sekvensielle tekstfiler kan ha sin opprinnelse langt tilbake i tid, noen så langt tilbake som til 1970-årene, og på denne tiden var systemene proprietære. Det vil si at et system som var utviklet på en IBM-maskin bare kunne brukes på IBM-maskiner, og de flate filene kunne ikke kopieres til en annen maskin. De flate sekvensielle tekstfilene var ikke laget i ASCII-kode, men i et eller annet format som datamaskinprodusenten hadde laget. Etter hvert kom standard ASCII-kode for de ulike tegn, og de fleste produsentene laget eksportprogrammer eller konverteringsprogrammer som overførte dataene fra det proprietære systemet til standard ASCII tekstfiler. Men fremdeles finnes det fagsystemer som bruker andre tegnsett, som for eksempel IBM stormaskiner som kan kjøre EBCDIC-kode.

Lagring av data på flate sekvensielle tekstfiler medfører at dataene er tilgjengelige og relativt lette å hente ut, men bearbeiding av datafiler og oppbygging av metadata vil kreve ressurser og kompetanse. Prosjektgruppen har produsert en konkret avlevering av et enkelt fagsystem fra en liten bedrift, for å opparbeide kompetanse på denne typen materiale.

# 2 Bearbeiding av flate sekvensielle tekstfiler

Selv om datafilene i utgangspunktet foreligger som flate sekvensielle tekstfiler, vil de i mange tilfeller kreve en viss grad av bearbeidelse for å tilpasse filene til arkivfaglige krav. Blant annet skal tekstfiler tilfredsstille avleveringsbestemmelsenes[1] krav til formater for langtidsbevaring og lagres som ren tekst – ISO 8859-1: 1998, Latin-1 og ISO 8859-4: 1998, Latin-4 for samiske tegn.

[1] Forskrift om utfyllende tekniske og arkivfaglige bestemmelser om behandling av offentlige arkiver, av 01.12.1999 nr. 1566, kp. VIII: Bestemmelser om elektronisk arkivmateriale som avleveres eller overføres som depositum til Arkivverket.

Ved mottak av en datafil fra et fagsystem vil det ofte være behov for å bearbeide datafilen litt, før den kan arkiveres. Hvis datafilen er dannet ved en eksport fra en database vil behovet for bearbeiding være lite, for da dannes datafilen først ved selve eksporten og kan påvirkes av miljøet rundt danningen til å bli slik arkivinstitusjonen ønsker det.

Hvis datafilen er hentet fra et system som bygger på dataregistre der hvert register er en ”flat” datafil, vil mer opprydding og bearbeiding være nødvendig. En slik datafil vil ofte inneholde slettede poster, der plassen i filen er fylt opp med slettetegn i stedet for de opprinnelige verdiene. I tillegg er slike datafiler ofte dannet i et ”gammelt” miljø der det er brukt 7-bits tegnkoder i stedet for 8-bits tegnkoder som er vanlig i Windows miljø. Det er vanligvis de norske spesialtegnene æ, ø og å som presenteres forskjellig i 7-bits og 8-bits miljø. Mest jobb med omdanning er hvis hele datafilen er dannet i et annet tegnmiljø enn ASCII, f.eks. EBCDIC som IBM brukte.

### 2.1 Endringer i postlayout over tid

Spesialsystem som blir laget for en bestemt bedrift/institusjon er ofte ikke statiske i oppbygging og innhold. Etter hvert som bedriftens/institusjonens behov og omgivelser forandrer seg, blir også spesialsystemene endret slik at de kan betjene bedriftens-/institusjonens endrede behov. Hvis data er organisert i flate tekstfiler med faste post- og feltlengder vil endring i postlengde kreve at alle programmer som bruker datafilene endres. Ofte blir dette forsøkt unngått ved at man legger inn ekstra plass ved design av systemet, og har muligheten til å endre defineringen av felter etter behov.

### 2.2 Fjerning av slettede poster

Slettede poster behøver ikke å være noe problem, de ligger bare som ”tomme” eller ”svarte” linjer ved en listing av registeret, men hvis det er mange slettede linjer vil lesbarheten minske og filen vil bruke unødvendig mye plass. Tidligere operativsystemer hadde kommandoer som kopierte dataene i filene, men hoppet over slettetegnene (COPY-/REFORMAT-kommandoer). Avleverende institusjon vil ofte ha datamaskiner med tilhørende operativsystemer tilgjengelig, og opprydding i datafilene kan skje før filene avleveres. Hvis det er slettede poster i en fil som skal avleveres/deponeres, finnes det verktøy som kan gjøre jobben. I programmeringsspråk som SAS, C eller C++ kan slike funksjoner programmeres og data på bitnivå manipuleres og kontrolleres.

### 2.3 Bytte tegnsett fra EBCDIC til ASCII

EBCDIC er en helt parallell tegnrepresentasjon med ASCII, bare mindre brukt. Det var miljøet rundt IBMs datamaskiner som brukte denne tegnrepresentasjonen, men IBM var helt dominerende i store etater og bedrifter slik at det ble laget mange systemer for disse maskinene. Konvertering fra EBCDIC til ASCII skjer enklest ved at det lages et program/en rutine i SAS som leser en EBCDIC-fil og skriver til en ny ASCII-fil. For Arkivverket er dette gjort ved mottak av datafiler fra Skattedirektoratet.

### 2.4 Skifte fra 7-bits til 8-bits tegnrepresentasjon

Når en tekst som hentes fram mangler æ, ø og å, er sannsynligheten stor for at teksten er dannet i et 7-bits miljø. Hvert tegn som brukes i datafilene er representert ved en binærkode som danner et tall (for eksempel er stor A lik 65, stor B lik 66 osv). Alle ”amerikanske” bokstaver og tegn er likt representert i 7-bits og 8-bits miljø. Derfor kan tekstene stort sett leses i begge miljøene, bare spesialtegnene endres. Konvertering av spesialtegn fra 7-bits til 8-bits kan gjøres uten noen store datakunnskaper.

En oppskrift kan være slik ved å bruke standard Microsoft WordPad:

1. Start opp programmet WordPad som følger med Windows.
2. Åpne den aktuelle filen
3. Finn en bokstav i filen som skal endres.
4. Kopier bokstaven (med Ctrl + C.)
5. Trykk på Rediger på kommandolinjen
6. Velg Erstatt i rullemenyen.
7. Kopier inn (CTRL + V) i feltet for ”Søk etter”.
8. Skriv inn den bokstav det skal endres til i ”Erstatt med” feltet.
9. Trykk på ”Erstatt alle”.

Punktene fra 3 til 9 må gjentas for hver bokstav som skal endres, i praksis seks ganger for store og små æ, ø og å. Når alle tegnene er byttet ut, må filen lagres og konverteringen er gjort.

# 3 Dokumentasjon av datafiler

## *3.1 Tekniske metadata*

For å muliggjøre gjenskaping og senere bruk av det elektroniske arkivmaterialet skal det medfølge en fullstendig struktur- og innholdsbeskrivelse av datafilene (metadata). Beskrivelsen skal omfatte datafilene og feltene i filene som avleveres eller deponeres, *ikke* strukturen i datafilene slik de forelå før bearbeiding.

Metadataene avleveres i elektronisk form, og skal være ren tekst i godkjent tegnsett. Fortrinnsvis skal den også benytte XML-syntaks, basert på syntaksreglene i en egen DTD (Document Type Definition) som Riksarkivaren har utarbeidet og publisert spesifikt for dette formålet: ADDMML - *Archives Data Description and Manipulation Markup Language*. Arkivverket har utviklet et eget verktøy, Arkadukt, for å produsere metadata som følger disse syntaksreglene. XML-filen kan så brukes for å behandle kildedataene som er lagret i de flate sekvensielle tekstfilene.

## *3.2 Annen systemdokumentasjon*

Jo mer systemdokumentasjon som medfølger, jo bedre bilde vil man kunne danne seg av konteksten datamaterialet er skapt i. Avleveringsbestemmelsene setter en rekke krav til medfølgende dokumentasjon, men for privat arkivmateriale vil man ikke kunne stille de samme kravene. Tekniske metadata er nødvendige for å kunne gjøre nytte av datafiler, og er derfor et absolutt krav, også for private arkiver. I tillegg bør man legge ved dokumentasjon som:

- System- og bruksdokumentasjon. Oversikt over det opprinnelige systemets overordnede struktur, en summarisk beskrivelse av systemets hovedfunksjoner, formålet med det systemet som har produsert arkivversjonen, bruken av det og den administrative sammenhengen det inngår i.
- Informasjon om avleveringen eller deponeringen. Navn på systemet som arkivversjonens datafiler kommer fra, tidsrommet som datafilene omfatter, antall poster i datafilene og annen tallfesting av omfang.

# 4 Hva kreves for å kunne bevare et fagsystem?

## 4.1 Teknologiske betingelser

For å kunne bevare informasjon fra fagsystemer med data organisert på flate sekvensielle tekstfiler kreves det at tekstfilene er tilgjengelige for den som skal produsere avleveringen. Det vil si at datafilene må kunne konverteres fra et proprietært dataformat til et standard ASCII-format, hvis det er nødvendig. Produksjon av tekniske metadata kan utføres i en enkel teksteditor, men bør utføres ved hjelp av Arkadukt.

## 4.2 Kompetanse

Bevaring av data fra fagsystemer av den typen som er omtalt i dette dokumentet, kan være en krevende oppgave. I mange tilfeller kreves det kompetanse om eldre operativsystemer og programvaremiljøer for å kunne hente ut datafiler fra fagsystemer. Videre vil bearbeiding av datafiler og produksjon av metadata kreve mer teknisk kompetanse enn avlevering av for eksempel tekstbehandlingsdokumenter eller elektronisk post (selv om disse kan ha andre utfordringer). Denne kompetansen må finnes hos arkivskaper eller i depotinstitusjon, eller man må benytte seg av ekstern ekspertise.