

Format- og Strukturkonverteringsprojektet

Artikel fra Nordisk Arkivnyt nr. 1, 2007

Af Jan Nepper-Christensen

Hvordan bevarer man egentlig gamle elektroniske arkivalier, hvis formater og strukturer snart hverken kan læses eller forstås længere? – Det spørgsmål stiller 14 ansatte i Format- og strukturkonverteringsprojektet (FSK) sig næsten dagligt i øjeblikket.

Arkivalierne

Rigsarkivet i København har igennem snart 30 år indsamlet elektroniske arkivalier fra den offentlige forvaltning i Danmark. Samlingen af ældre arkivalier, dvs. afleveret før 2000 omfatter 1253 såkaldte arkiveringsversioner med bevaringsværdige data. Det drejer sig om journaler og registre fra 1960'erne til ca. 1998. Det er eksempelvis CPR-registeret tilbage fra dets oprettelse i 1968, Undervisningsministeriets Elev, skole og klassestatistik fra 1963 og Statsskattedirektoratets slutligning fra 1970 for blot at nævne nogle af de ældste registre. Derudover omfatter arkiveringsversionerne både stort og småt lige fra domstolenes journaler, BBR-registeret, Københavns Universitets studieregister, private forskningsdata til Miljødata fra 1990'erne. Data fylder i alt næsten en terabyte (1024 gigabyte), og de ligger sikkert bevaret på bevaringsmedier, som pt. er harddiske i RAID's og CD-R.

Alle disse arkiveringsversioner skal nu format- og strukturkonverteres, således at de kan komme til at overholde moderne standarder. De vil efter konvertering komme til at overholde samme bevaringsformat som alle de nyere arkivalier afleveret senere end 2000. Bevaringsformatet specificeres i *"bekendtgørelse nr. 342 af 2004 om arkiveringsversioner af bevaringsværdige data fra elektroniske arkivsystemer"*.

Formålet med konverteringen er at muliggøre en effektiv og sikker bevaring nu og fremover. Derudover er konvertering en forudsætning for en standardiseret tilgængeliggørelse af disse arkivalier.

Konvertering af tegnsæt og hierarkiske databaser

Mange af arkivalierne er altså afleveret til Rigsarkivet i 70'erne og 80'erne og de ligger gemt ned i tegnsæt, som i dag ikke benyttes længere. Læser man det ind i et moderne tekstbehandlingsprogram vises kun noget volapyk på skærmen, der mest af alt ligner noget man kan finde i taleboblen over Donald Duck, når han er rasende. Det er fuldstændig ulæseligt! Se illustrationens højre side, hvor den samme post i Københavns Havns skibsregister vises både i gammelt tegnsæt (den ulæselige) og i nyt tegnsæt.

NR.	NAVN	TYPE	LÆNGDE	INDHOLD eller fejlbeskrivelse
1	AHH96ID	STRING_001	2	ùö
2	DATOMM	NUM_001	2	ðñ
3	DATOAA	NUM_001	2	øö
4	EXPNNR	NUM_001	5	ðñððð
5	SAEJRUTE	STRING_001	1	@
6	KUNRPCAK	NUM_009	3	□□,
7	MOMSKODE	STRING_001	1	@
8	FILLER_1	STRING_001	6	ðððñøö
9	DISTRIKT	NUM_001	1	ó
10	KAJ	NUM_001	3	øó+
11	GODSMGD	NUM_001	5	ððñðð
12	VAREKODE	NUM_001	3	ð+ð
13	IUKODE	NUM_001	3	ðóö
14	LOLAKODE	NUM_001	2	ññ
15	FILLER_2	STRING_SPACE	2	@@
16	NUM8DEC2	REAL 004D02	8	ðñððøððð

NR.	NAVN	TYPE	LÆNGDE	INDHOLD eller fejlbeskrivelse
1	AHH96ID	STRING_001	2	96
2	DATOMM	NUM_001	2	01
3	DATOAA	NUM_001	2	85
4	EXPNNR	NUM_001	5	21062
5	SAEJRUTE	STRING_001	1	
6	KUNRPCAK	NUM_009	3	13142
7	MOMSKODE	STRING_001	1	
8	FILLER_1	STRING_001	6	020185
9	DISTRIKT	NUM_001	1	4
10	KAJ	NUM_001	3	837
11	GODSMGD	NUM_001	5	02120
12	VAREKODE	NUM_001	3	270
13	IUKODE	NUM_001	3	236
14	LOLAKODE	NUM_001	2	11
15	FILLER_2	STRING_SPACE	2	
16	NUM8DEC2	REAL 004D02	8	010282,00

Knap halvdelen af data har ikke kun gammeldags tegnsæt – de kommer tilmed fra gammeldags databasesystemer. De ældre systemer gemte data i hierarkisk form – eller sekventielle datastrukturer som nogle foretrækker at kalde det. Hierarkiske databaser er forløberen for relationelle databaser, som er de mest udbredte i dag. De gamle hierarkiske strukturer konverteres til relationel form.

Konverteringsprincipper

Imidlertid kan man jo ikke konvertere uden at ændre, det siger sig selv. Men hovedprincippet må nødvendigvis være, at ændre så lidt som muligt, når der er tale om arkivalier. Men det er ikke altid nogen nem sag. Der dukker konstant metodiske problemstillinger op undervejs. Fx når der i data findes ulovlige tegn, som tydeligvis ikke betyder det samme hver gang. I de tilfælde er en meningsfuld tolkning af data umulig, og tegnene konverteres simpelthen til et omvendt spørgsmålstegn, for på denne måde at signalere overfor den fremtidige bruger at datas betydning er ukendt.

Princippet med relationelle databaser er, at felter i forskellige tabeller skal relateres til hinanden ved hjælp af nøgler, og værdierne i det ene nøglefelt skulle gerne være en delmængde af værdierne i det andet. Men det er langt fra altid tilfældet, selvom det burde være det. Fremmednøglen indhold passer simpelthen ikke (fuldstændigt) med primærnøglen værdier – den referentielle integritet er overtrådt. Det er vi nødt til at acceptere til en vis grad, samtidig med, at der foretages en vurdering af, om en relation, hvor den referentielle integritet er overtrådt, virkelig ER en autentisk relation.

Digitalisering af lyd- og videobånd

Udover alle de gamle datas konvertering skal format- og strukturkonverteringsprojektet også digitalisere Statens Arkivers lyd- og videomateriale. Det drejer sig om at få overspillet lakplader, kassettebånd, spolebånd og en række forskellige typer videobånd til digitalt format. Det åbner mulighed for, at dette materiale også kan gemmes som arkiveringsversioner og dermed kan håndteres og bevares efter samme principper som det øvrige digitale materiale. Statens Arkiver ligger også inde med en række filmruller, men de digitaliseres ikke. Det er at foretrække, at opbevare de meget holdbare filmruller i magasiner med det rette klima frem for at gennemføre voldsomt dyre digitaliseringer.

Projektstatus

Men hvad skal der så ske med alle disse gamle data? Jo, det første år af projektet er gået med en række forberedelser. Data lå ikke struktureret ens og var ikke registret i arkivdatabase, så det skulle først ske. For at konverteringen overhovedet kunne foregå, var det desuden nødvendigt med et konverteringsprogram, et program, som kan konvertere de gamle data til nyt bevaringsformat. Det har krævet megen tid at programmere og udtænke et koncept, som kan håndtere det næsten uendelige antal variationer, som strukturer og formater kan forekomme i. Der udvikledes derfor et (xml-baseret) digitalt beskrivelsesformat, som er blevet omdrejningspunktet i hverdagen for de ca. 9 medarbejdere, som lige nu konverterer arkivalierne. Derudover er der foregået en masse andre forberedelser blandt andet skanning af alle datas tilhørende papirdokumentation og meget andet.

I slutningen af 2006 stod vi ved en skillevej. Fra at have tænkt, udviklet og forberedt os i 1½ år var det tid til at få konverteret de mange data. Inden årets udgang blev de første 15 % af konverteret, som også lovet i resultatkontrakten med Kulturministeriet, som har bevilget de godt 15 mio. kr., som projektet er budgetteret til.

Målet er, at vi ved slutningen af 2008 har nået vores mål om at have konverteret alle ældre elektroniske arkivalier til den gældende standard, og at alle lyd- og videomedier ligeledes er digitaliseret. Dermed skulle alle e-arkivalier, samt lyd og videoarkivalier være sikret bevaring et godt stykke tid ud i fremtiden.

Forfatteren er cand. mag i historie og merkonom i IT og ledelse, samt projektleder for Format- og strukturkonverteringsprojektet.