

Kriterier for Statens Arkivers valg og fravalg af filformater i forbindelse med langtidsbevaring af arkivalier

I det følgende er der mulighed for at læse om kriterierne for Statens Arkivers valg af tilladte filformater til aflevering af elektroniske arkivsystemer.

Grundlæggende strategi

En af Statens Arkivers opgaver er at sikre at e-arkivalier kan læses af eftertiden, hvilket indebærer at vi dels skal kunne modtage e-arkivalier, men også opbevare og vedligeholde disse. Der findes i princippet 2 strategier for at kunne opfylde denne opgave:

1. Kendskab og håndtering af alle filformater anvendt i den offentlige forvaltning.
2. Udpegning af få, men tilstrækkelig mange, veldefinerede og velkendte formater, som myndighederne på afleveringstidspunktet kan benytte, uden at vi af den grund virker hæmmende på den udvikling, som foregår i IT-verdenen.

Strategi 1 er fravalgt af den grund, at udviklingen indenfor IT-verdenen går så hurtig, at håndteringen af flere hundrede filformater af ressourcemæssige og licensmæssige årsager ikke lader sig gøre.

Strategi 2 har den fordel, at den er overskuelig og ressourcemæssig realistisk, men har samtidig den ulempe, at det kan være overordentlig vanskelig at udpege velegnede formater. Heldigvis er det ved at gå op for producenter af dokumenthåndteringssystemer, at de selv får et vedligeholdelsesproblem. Så tendensen går mod anvendelse af mindre komplekse filformater uden for mange indbyggede funktioner.

Grundlæggende krav til formater som Statens Arkiver tilstræber overholdt

- Formatet skal være standardiseret (ISO, ANSI eller lign.) eller som minimum velbeskrevet (f.eks. TIFF).
- Formatet skal være bredt understøttet.
- Formatet skal være platform-uafhængigt og åbent (ikke behæftet med licens eller lign.).
- Formatet må ikke være tabsgivende på en sådan måde at kvaliteten af data forringes.
- Formatet skal have en lang forventet levetid.
- Formatet skal kunne konverteres til nyere kommende formater, dvs. at formatet ikke må indeholde specielle funktioner eller lign., som relaterer sig til bestemte operativsystemer eller programmer (OLE-Objekter - f.eks. @gule sedler@).
- Formatets fremtrædelse skal være uafhængig af fonte mv. hvis fremtrædelsen er af betydning, (i tekst-versioner af et arkivalie som benyttes til søgning, er fremtrædelse uden betydning).

Forskellig typer data stiller forskellige krav til filformat

Dokumenter

Her skelnes mellem anvendelse:

1. Til tekst som man ønsker at kunne søge i, kan ASCII eller ANSI-tekstformatet benyttes. Grafik og formateringer (fed, kursiv mv.) understøttes ikke, hvorfor dokumentets oprindelige udseende ikke kan bibeholdes.
2. Til formater som understøtter formatering og indlejret grafik mv. vil man blive nødt til at vælge en af de formater som findes og benyttes på det nuværende marked (f.eks. Word eller WordPerfect). Hvis

dokumenternes oprindelige udsende skal bevares, er man nødt til at bevare de fonte som er benyttet i dokumenterne hvilket ikke er realistisk fordi de afhænger af operativsystem mv.

3. Til ikke søgbare/låste filformater kan bitmap benyttes. Dokumentets udsende bevares, som da det blev skabt og det er ikke muligt at redigere dokumentet, hvilket må siges at være en stor sikkerhedsmæssig fordel. Ulempen er at det ikke umiddelbart er muligt at søge i dokumentet, men ved at benytte OCR kan der skabes en tekstversion af det oprindelige dokument som er søgbar.

Hypertekst

Med hypertekst menes Windows hjælpefiler og filer af SGML-typen (HTML/XML), hvori det er muligt at definere links og/eller bogmærker.

Bitmap/foto

Et elektronisk foto består af umådelig mange punkter (pixels), og afhængig af antallet af farver som disse punkter skal kunne antage, fylder hvert punkt hvad der svarer til en eller flere tegn. En dokumentside med ca. 2000 tegn vil i en bitmapudgave fylde, hvad der svarer til flere millioner tegn! Løsningen på dette problem søges løst via kompressionsalgoritmer der, i stedet for at gemme de enkelte punkter, gemmer hele områder. Hermed kan filstørrelsen nedsættes kraftigt. Filtypen JPG går skridtet videre og justerer på farvenuancer, således at områder hvor farvenuancerne er næsten ens ... bliver ens!

Original



På afstand kan det være svært at se forskel, men når man zoomer ind på detaljer (slipseknuden !) bliver forskellen tydelig (JPG billedet er opbygget af små firkanter).

Ikke tabsgivende kompression



Tabsgivende kompression (JPG)



JPG fylder meget lidt på harddisken, men er altså tabsgivende. I bestemte situationer er dette en stor fordel, f.eks. til brug på Internettet hvor kvaliteten af billeder spiller en mindre rolle i forhold til den mængde data, som skal flyttes over netværk.

Lyd

Hvad er kvalitet, og hvad er @egentlig data indhold@ ? En telefonsamtale og en klaverkoncert stiller meget forskellige krav til kvalitet og dermed måden, hvorpå data kan gemmes. Det er ikke nok, at vi kan genkende musikken, og på den anden side er der ingen grund til at gemme en telefonsamtale i CD kvalitet. I det ene tilfælde er det indholdet af telefonsamtalen, og i det andet er det summen af samtlige nuancer, som er det egentlige dataindhold. Lyd fylder meget, og når Statens Arkiver skal håndtere langtidsopbevaring af lyd, bør vi skelne mellem typer af lyd. Tabsgivende - men effektiv kompression - bør evt. tillades i de situationer, hvor kravet til kvalitet er lav.

Levende billeder

Levende billeder er på mange måder sammenlignelig med lyd. Måden at gemme lyd på afhænger af formålet med data. Hvis der er tale om en videokonference, hvor genkendelighed er nok (det som siges og hvem som siger hvad), så er kvaliteten af mindre betydning. Hvis der derimod er tale om optagelser, hvor det billedlige indhold skal være nuanceret, skarpt og så naturtro som muligt, så er kvaliteten af stor betydning. Levende billeder består af mange billeder, der som ofte kan være af en forholdsvis dårlig kvalitet, hvis man fokuserer på det enkelte billede, men af en rimelig kvalitet, når de ses som helhed (et tv-stil-billede fra en videooptagelse er af ringe kvalitet). Tabsgivende, men effektiv kompression, bør evt. tillades i de situationer, hvor kravet til kvalitet er lav.

CAD Konstruktionstegninger og lign. gemmes i såkaldte vektoriserede grafikformater. I stedet for at gemme de enkelte punkter i en linie eller en kurve, så gemmes x,y koordinater eller en matematisk formel. CAD tegninger rummer ofte flere lag og kan skaleres frit. Der findes flere velbeskrevne formater, som har været eller fortsat er meget udbredte, men den teknologiske udvikling og udvikling indenfor CAD-verdenen har ført til nye formater, som f.eks. gemmer informationer i databaseform, og som derfor er mindre tilgængelige og langt mere komplekse end de foregående formater. Problemet er, at det vil være besværligt, måske umuligt, at konvertere nyere tegninger til ældre velkendte standardiserede og uforanderlige formater.

Regneark

Et regneark vil oftest bestå af mere end blot rækker og kolonner med enkelte funktioner så som summen af tal, dato etc. Med de nuværende benyttede regnearksformater, er det muligt at gemme komplekse statistiske funktioner, Pivot tabeller mv. Problemet er, at det ikke vil være muligt at konvertere til ældre velkendte standardiserede og uforanderlige formater.