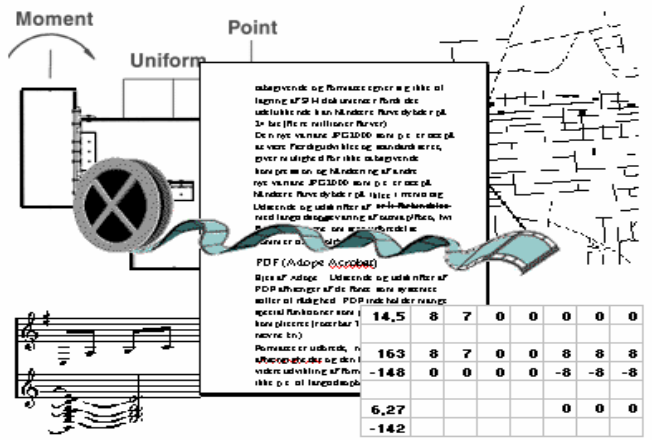
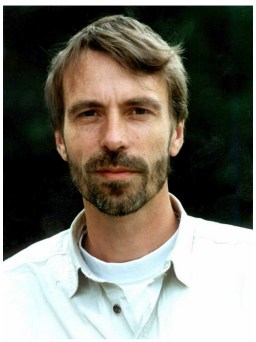
Filformater ...

Baggrunden for Statens Arkivers valg og fravalg af filformater til brug for langtids-opbevaring af arkivalier, samt gennemgang af de mest udbredte filformater.



Af René Mittå Olsen
Systemudvikler, Statens Arkivers IT-afdeling



Udviklingen inden for elektronisk sags- og dokumenthåndtering er i en rivende udvikling, men hvordan står det til med langtidsopbevaringen af disse informationer?
Hvorfor er Statens Arkiver tilbageholdende med at tillade diverse filformater i forbindelse med elektroniske afleveringer?
Hvorfor findes der så mange forskellige filformater til håndtering af tekst, grafik, lyd, regneark, video mv., og hvad er de forskellige formaters styrke og svaghed?
Alle snakker om XML ... men hvad er XML egentlig?

Indledning

Overgangen fra papirbaseret til digital sagsbehandling er i fuld gang.
Elektroniske sags- og dokument – håndteringssystemer (herefter ESDHsystem) er ved at vinde indpas i den offentlige forvaltning.

Sager som før bestod af håndskrevne eller udprintede notater, afskrifter af møder og telefonsamtaler (evt. krydret med tegninger og fotografier), er i dag på vej mod fuldstændig digitalisering.
Et elektronisk arkiv kan i dag bestå af digitale dokumenter, tekst, billeder, regneark, gule sedler, GIS, lyd, video etc. Disse digitale dokumenttyper kan være mere eller mindre integrerede i hinanden, og alle er de ét museklik fra redigering eller visning/afspilning.

Fælles for disse dokumenttyper er, at de skal kunne gemmes som filer, og at de oftest repræsenterer viden som vi ønsker skal leve længe.

Grundlæggende problematik

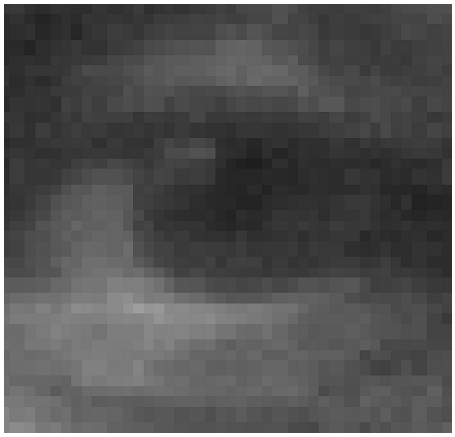
Det går hurtigt!
Leverandørerne af Officepakker vælger ofte, af konkurrencemæssige årsager, at lægge vægt på programmernes funktionalitet og at opretholde et hurtigt udviklingstempo (samt at løse de problemer et hurtigt udviklingstempo afstedkommer).
På området vedr. ESDHsystemer er der også ved at komme gang i udviklingen, og det er ikke ualmindeligt, at et ESDHsystem er i stand til at håndtere 30-60 forskellige filformater. Antallet af filformater benyttes ofte som en konkurrenceparameter over for kunden, men er der tænkt på:

- at myndigheden, som køber systemet, er forpligtiget til at sikre informationernes overlevelse?

- at disse informationer, for at overleve over tid, løbende skal konverteres til nye ”levende” formater?
- at mange formater er fejlbehæftede og sårbare overfor konvertering?
- at kunden måske en dag ønsker ESDHsystemet konverteret til et andet system (evt. fra en anden leverandør)?

Et eksempel

For at illustrere problemet vil jeg dykke lidt ned i en gruppe af disse formater, nemlig formater til håndtering af billeder (herefter kaldt bitmap).



Et bitmap billede består af en masse små punkter kaldet pixels.
Bitmap filerne, hvori billederne gemmes, er grundlæggende opbygget af en filheader (med oplysninger om billedets højde, bredde, antal farver, kompressions metode, copyright etc.) efterfulgt af data (bestående af en lang række tal som hver især beskriver én pixel eller en del af én pixel).
Selvom alle bitmap filformater stort set er opbygget på samme måde, så findes der alligevel flere hundrede forskellige. Af de mest kendte kan nævnes TIFF, JPG, GIF, BMP, TGA, EPS og PCX.
Forskellen på disse formater er ofte minimale og er opstået på grund af copyright, forsøg på at fastholde brugerne til de enkelte programmer samt i et vist omfang ”genopfindelse af den dybe tallerken”. Hertil kommer at mange er dårligt beskrevet (ikke standardiserede), og derfor findes i flere varianter, som desuden håndteres forskelligt fra program til program.

Af disse mange bitmapformater er der kun få som fuldt ud opfylder de krav man bør stille, når der er tale om opbevaring af elektroniske arkivalier. En tendens som desværre også er gældende for andre typer af dokumentformater.

Grundlæggende krav til filformater 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 platformuafhængigt og åbent (ikke proprietært og 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. Det vil sige, at formatet ikke må indeholde specielle funktioner eller lign. som relaterer sig til bestemte operativsystemer eller programmer (OLE Objekter - f.eks. ’gule sedler’).
- Formatets fremtræden skal være uafhængig af fonte mv., hvis dette er af betydning (i tekstversioner af et arkivalie som benyttes til søgning, er dokumentets fremtræden uden betydning).

Statens Arkivers strategi for valg/fravalg af filformater

Statens Arkiver hilser tidens nye digitale muligheder velkomne. Tænk hvad det betyder i sparet magasinplads, hurtig adgang til de elektroniske arkivalier og dermed forbedret service over for Statens Arkivers private såvel som offentlige kunder.
Men .... Statens Arkivers primære opgave 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.
Statens Arkiver har valgt den såkaldte konverteringsstrategi for at sikre at

elektroniske arkivalier også kan læses og forstås i fremtiden. Det er derfor nødvendigt at udpege få (men tilstrækkelig mange) veldefinerede og velkendte formater, som myndighederne på afleveringstidspunktet kan benytte, og som Statens Arkiver, med indiskutabel sikkerhed ved kan konverteres til fremtidige formater.

Forskellig typer data stiller forskellige krav til filformat

Det skal bemærkes at det ikke er Statens Arkivers krav som her diskuteres, men de forskellige dokumenttypers styrke og svagheder generelt.

Tekstdokumenter

Her skelnes mellem anvendelse:

- Til tekst der skal kunne søges i, kan ASCHII eller ANSI tekstformatet benyttes. Grafik og formateringer (fed, kursiv mv.) understøttes ikke, hvorfor dokumentets oprindelige udseende ikke kan bibeholdes.

- 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 udseende skal bevares, er man også nødt til at bevare de fonte som er benyttet i dokumenterne, fordi disse fonte afhænger af operativsystem mv.

- Til ikke søgbare/låste filformater kan bitmap benyttes. Dokumentets udseende 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 hjælp af OCR (Optical Character Recognition) kan der skabes en søgbar tekstversion af det oprindelige dokument.

Hypertekst

Hypertekst er en funktionalitet, som gør det muligt at definere links og/eller bogmærker, som gør at brugeren kan springe mellem interne dele af et dokument, og/eller til eksterne dokumenter.
Windows hjælpefiler, filer af SGMLtypen (HTML/XML), men også almindelige tekstdokumenter, kan indeholde både statiske og/eller dynamiske links. Her er det vigtigt at være opmærksom på, at det de dokumenter som et dynamiske link peger på, kan blive ændret eller ligefrem slettet over tid. En af løsningerne kan være at sikre sit dokument ved at hente kopier af de dokumenter der linkes til, og gemme disse kopier sammen med selve dokumentet.

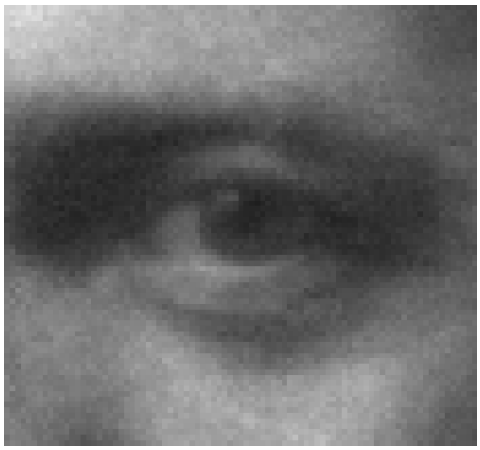
Bitmap/foto

Et elektronisk foto består, som nævnt i det foregående, 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 side i et tekstdokument bestående af ca. 2000 tegn, vil i en bitmapudgave fylde hvad der svarer til flere millioner tegn.

Dette problem søges løst via kompressionsalgoritmer der - i stedet for at gemme de enkelte punkter - gemmer større eller mindre områder (f.eks. pixellinier). Hermed kan filstørrelsen nedsættes kraftigt. Filtypen JPG går skridtet videre og justerer desuden på farvenuancer, således at områder hvor nuancerne er næsten ens ... omdannes til at blive helt ens.

Originalt billede

På afstand kan det være svært at se forskel, men når man zoomer ind på detaljer bliver forskellen tydelig.



Udsnit fra originalbillede gemt som TIFF.

Efter konvertering til JPG.
Bemærk de små firkanter som et JPG billedet er opbygget af.

JPG fylder meget lidt på harddisken, men er altså tabsgivende. I visse situationer er dette en stor fordel (f.eks. til brug på Internettet hvor kvaliteten af et billede spiller en mindre rolle i forhold til den mængde data som skal flyttes over netværk).
Det er muligt at tilpasse kompressions niveauet således at datatabet bliver mindre, men så vokser filstørrelsen, og dermed behovet for fysisk lagerplads tilsvarende.

Lyd
Hvad er kvalitet, og hvad er egentligt 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 CDkvalitet. I det ene tilfælde er det summen af samtlige nuancer i musikken som er det egentlige dataindhold, og i det andet er det genkendelsen af telefonsamtalens ord.
Lyd fylder meget, og man bør derfor skelne mellem typer af lyd og dermed behovet for kvalitetsniveau, således at tabsgivende - men effektiv - kompression kan 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, er kvaliteten af stor betydning.

Levende billeder består af mange billeder, der er af en forholdsvis dårlig kvalitet hvis man fokuserer på det enkelte billede, men af en rimelig kvalitet når billederne ses som helhed (et TV stillbillede fra en videooptagelse må siges at være af ringe kvalitet).
Levende billeder fylder ekstremt meget, og man bliver derfor nødt til at benytte en eller anden form for effektiv kompression når disse skal lagres elektronisk.
Det er muligt at tilpasse omfanget af kompressionen (læs datatabet) således at det er tilpasset ønsket om kvalitet.

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. CADtegninger rummer ofte flere lag og kan skaleres frit.
Den teknologiske udvikling og udvikling inden for CADverdenen har ført til nye formater som er langt mere komplekse end de ældre foregående formater, og flertallet

af disse er proprietære. Problemet er at det med tiden kan blive besværligt (måske umuligt) at konvertere nyere tegninger til ældre, velkendte, standardiserede og uforanderlige formater.

GIS
GIS er en forkortelse af Geografisk Informations System. Med GIS skal forstås elektroniske systemer til opbevaring og visualisering af geografiske kort.
GISbilleder gemmes primært som vektoriserede linier og kurver, og med de informationer som knytter sig til kortets forskellige områder.
GIS er på mange måder sammenlignelig med CADfiler.

Regneark
Et regneark vil oftest bestå af mere end blot rækker og kolonner med simple funktioner såsom summen af tal, dato etc.
I nyere regnearkformater er det muligt at gemme komplekse statistiske funktioner (Pivottabeller o.lign.).
Problemet er at det ikke vil være muligt at konvertere til ældre, velkendte, standardiserede og uforanderlige formater.
Noget tyder på at løsningen på problemet med lagring af regneark skal findes via ekstensions til XML (se afsnit om XML).

Metaformater
Med metaformater menes formater som er i stand til at håndtere/indeholde mange forskellige typer af data.
Microsoft Word formatet DOC, Corels WPD (Wordperfect) og Adobes PDF format er eksempler på metaformater som udover tekst, kan rumme billeder, videosekvenser, regneark, hyperlink etc.
Problemet med de nævnte formater er at de alle er proprietære formater i hastig udvikling, hvilket indebærer at der ikke er nogen garanti for at formatet vedligeholdes, endsige eksisterer, i tiden frem. Der er desuden tale om komplekse formater, som det kan være vanskeligt at konvertere til andre formater.

Gennemgang af et udvalg af filformater

I det følgende beskrives et udpluk af filformater udvalgt på grund af udbredelse og myndigheders ønsker om at måtte benytte dem i forbindelse med aflevering af elektroniske arkivalier.
Statens Arkivers vil - for hvert enkelt filformat – kommentere formatets egnethed i forhold til langtidsopbevaring af arkivalier, og dermed begrunde Statens Arkivers nuværende tilvalg og især fravalg af det pågældende format som afleveringsformat.
Der er lagt en del vægt på en beskrivelse af formatet XML, fordi meget tyder på at XML vil komme til at spille en markant rolle som fremtidens bærer af informationer af enhver art.

TIFF

TIFF blev oprindeligt udviklet og specificeret i 1986 af Aldus (Aldus blev senere opkøbt af Adobe). Sidste revision - før Aldus stoppede med at udvikle på formatet - var TIFF revision 6.0 fra 3. juni 1992.

TIFF er overordentlig udbredt og understøttes af stort set alle skannerprogrammer og programmer til håndtering af grafik og dokumenter.
En af styrkerne ved TIFF er at formatet er meget robust og enkelt opbygget, men samtidigt rimeligt avanceret.
Statens Arkiver tager udgangspunkt i TIFF revision 6.0, som specificerer det basale format (det som enhver TIFF viewer skal kunne håndtere) og registrerede ekstensions (funktionalitet og features som en TIFF viewer evt. har valgt at kunne håndtere).

Det er ifølge TIFF revision 6.0 muligt at benytte følgende kompressionstyper:

- Packbit kompression
- CIT/TSS Group 3 og 4 kompression
- LZW kompression
- JPEG kompression

Statens Arkiver har valgt at benytte LZW, fordi den er effektiv til kompression af farve/gråtone bitmap, og CCIT/TSS gruppe 4 kompression algoritmen til S/H bitmap, fordi der er tale om en uhyre effektiv kompression.
LZW er licensbehæftet (UNISYS), men udelukkende når der er tale om udvikling af kommerciel software (Statens Arkiver har af samme grund tilladelse til frit at benytte og distribuere software som benytter LZW, til myndigheder og leverandører) .

Statens Arkiver har undladt at bruge JPEG kompression fordi algoritmen er tabsgivende, og undladt brugen af Packbit kompression, eftersom den ikke er effektiv nok.

Selvom det er muligt at definere en multiplepage color TIFF filtype som overholder Baseline TIFF 6.0 specifikationerne, så har Statens Arkiver valgt ikke at benytte denne mulighed, fordi disse (proprietære) varianter ikke er udbredte og derfor endnu ikke understøttes i tilstrækkelig omfang.

DOC (WORD)

DOC er et Microsoft dokumentformat som er meget udbredt. Desværre ændres formatet fra programrevision til programrevision, senest sandsynligvis på grund af et ønske om en tættere forbindelse til Internettets nuværende HTML format og det kommende XML format.
Word formatet må siges at være et kompliceret format under fortsat udvikling, og formatet er ikke standardiseret, og må derfor anses som værende uegnet til langtidsopbevaring af arkivalier.

JPG

JPG benyttes til lagring af billeder, og udvikles af en gruppe (Joint Photographic Experts Group) under standardiseringsrådene ISO/IEC.
JPG er et filformat, som benytter en kompressionsmetode kaldet JPEG. Filformatet JPG og kompressionsmetoden JPEG er nærmest synonyme størrelser, men andre filformater end JPG, kan sagtens benytte JPEG kompressionsmetoden.
Selvom JPG er meget udbredt og understøttes af alle grafikprogrammer og Internet browsere, egner formatet sig ikke til langtidsopbevaring af arkivalier. Kompressionsalgoritmen som anvendes er tabsgivende og formatet egner sig ikke til lagring af S/H dokumenter fordi det kun kan håndtere farvedybder på 24 bit (flere millioner farver).

En ny JPG standard ”JPG2000” som for nyligt blev færdigudviklet, giver mulighed for ikke tabsgivende kompression og håndtering af andre farvedybder end 24 bit (16,8 millioner samtidige farver).
”JPG2000” vil ud fra et arkivmæssigt synspunkt sandsynligvis være velegnet til langtidsopbevaring af bitmap/foto, hvis formodningerne om stor udbredelse kommer til at holde stik.

PDF

PDF ejes og udvikles af firmaet Adobe, og benyttes som internt format i firmaets produktserie Acrobat.
Adobes PDF reader er gratis og er understøttet på alle operativsystemer (platform uafhængig), hvilket har været med til at sikre formatet stor udbredelse som udvekslingformat på Internettet.
PDF indeholder mange specialfunktioner, og de mange plugins (f.eks. roterbar 3D grafik for bare at nævne én) er med til at komplicere formatet og dermed mulighederne for fejl.
Udseende og udskrifter af PDF kan skifte afhængig af hvilke fonte et system stiller til rådighed.
PDF er udbredt, men pga. af fontafhængigheder og den fortsatte videreudvikling af formatet, skønnes PDF ikke at være velegnet til langtidsopbevaring af arkivalier.

AVI

AVI (Video for Windows) benyttes til at håndtere animationer og levende billeder.
AVI formatet er ikke standardiseret, men er velbeskrevet (Microsoft de facto standard).
Det er meget udbredt men på kraftig retur.
Selvom formatet ikke direkte benytter tabsgivende kompression, nedsættes billedkvaliteten i forbindelse med produktionen (downsampling) således at der reelt er tale om tab af kvalitet. AVI benytter WAV som lyddel (se beskrivelse af WAV).
Potentiel mulig som format til langtidsopbevaring pga. udbredelse og fordi der ikke mere udvikles på formatet, men AVI er ikke egnet til at håndtere levende billeder i TV kvantitet/kvalitet.

MPEG

MPEG benyttes til levende billeder og udvikles af en gruppe (Moving Picture Experts Group) under standardiseringsrådene ISO/IEC.
MPEG version 1,2 og 4 anvender alle en kompressionsalgoritme som ligner den som benyttes af bitmapformatet JPG.

MPEG 1,2 og 4 kompressionsalgoritmen giver datatab, og hvis man ser på det enkelte billede (stillbillede), vil datatabet kunne ses med det blotte øje, men når der skiftes mellem billederne (filmen afspilles) bemærkes dette ikke.
MPEG findes i flere versioner hvoraf MPEG 2 er den mest udbredte. MPEG 4 kan i princippet det samme som MPEG 2, men er blevet forbedret og udvidet på en lang række punkter. MPEG 4 er endnu ikke slået igennem som afløser for den mere enkle version 2.
MPEG 2 skønnes velegnet som format til håndtering af video pga. udbredelse og fordi der er tale om et velbeskrevet, standardiseret format som er i stand til at håndtere levende billeder i TV kvantitet/kvalitet på linie med DVD film.
Statens Arkiver er i skrivende stund ved at undersøge formatet nærmere med tanke på valg af MPEG 2 som format til langtidsopbevaring af levende billeder.

WAVE

Benyttes til lagring af lyd, og er meget udbredt, men på retur. Formatet er ikke standardiseret, men velbeskrevet (de facto standard) og i stand til at lagre lyd i CD kvalitet.
WAVE findes i mange varianter hvilket stiller store krav til den software som skal håndtere formatet.
Potentiel mulig som format til langtidsopbevaring af lyd pga.. udbredelse og fordi der ikke længere udvikles på formatet.

MP3

Benyttes til håndtering af lyd og er meget udbredt.
MP3 benytter en tabsgivende kompressionsalgoritme MPEG 1 (JPEG lign.), men er i stand til at gengive lyd i en kvalitet som nærmer sig CD kvalitet.
Der udvikles fortsat nye versioner af formatet, men formatet er udbredt og standardiseret og skønnes derfor velegnet som format for lyd.
Statens Arkiver er i skrivende stund ved at undersøge formatet nærmere med tanke på valg af MP3 til langtidsopbevaring af lyd.

XLS

XLS benyttes til lagring og udveksling af regneark i Microsoft programmet Excel.
Formatet er proprietært, ikke standardiseret og er under fortsat udvikling.
Formatet skønnes derfor uegnet til langtidsopbevaring af arkivalier.

DSF/FLA

DSFL er en forkortelse for Dansk Selskab for Fotogrammetri og Landmåling.
Selskabet har siden 1983 udviklet og vedligeholdt et dansk format til lagring og udveksling af GIS data kaldt DSFL.
Formatet er ikke kommercielt, standardiseret, og velegnet til langtidsopbevaring af GIS informationer.

DXF

DXF er et udvekslingsformat som er udviklet, og vedligeholdes af firmaet AutoDesk. Formatet er meget udbredt, velbeskrevet og frit tilgængeligt.
En af fordelene ved formatet er at det lagres som simpel ACHII tekst, og dermed er umiddelbart læseligt. En mindre ulempe ved anvendelsen af ACHII, er at det dermed optager mere plads end formater som lagres binært.
Selvom der fortsat tilføjes nye objekttyper til formatet (udvidelser kræver opdateret software), så er formatet velegnet til langtidsopbevaring af eksisterende og fremtidige CAD tegninger.

DWG

DWG er et filformat som er udviklet og vedligeholdes af firmaet AutoDesk.
Formatet er meget udbredt, og forsøges gjort frit tilgængeligt af OpenDWG Alliance (sammenslutning af CADbrugere).
Potentielt egnet til opbevaring af CADtegninger, specielt hvis formatet gøres frit.

HTML

HTML (Hyper Text Markup Language) egner sig til arkivering af Internet sider, men formatet er opbygget på en sådan måde at udseende i høj grad afhænger af det program som viser filens indhold (browseren), installerede fonte og grafikformaterne GIF og JPG.
HTML formatet er i et vist omfang fejlbehæftet og inkonsistent, men på grund

af sin store udbredelse potentiel mulig som lagringsformat.

XML

XML (eXtensible Markup Language) er et sæt af regler for opmærkning af simpel ACHIItekst således at denne tekst opdeles i veldefinerede logiske elementer, f.eks. TITEL, FORFATTER , TABELNAVN, DATO o.lign.
Formålet med denne opmærkning er at opbygge veldefinerede strukturer af data som efterfølgende kan benyttes til snart sagt ethvert formål f.eks. intelligente søgninger i dokumenter, definition af tabellerne i en database (metadata) eller hjemmesider/dokumenter på Internettet hvor brugeren har mulighed for at se forskellige udsnit af disse (f.eks. resume, fuld tekst, indholdsfortegnelse etc.).

XML definerer altså en struktur i en tekst, men fortæller intet om hvad denne struktur skal bruges til, og heller intet om hvordan denne tekst kan eller skal præsenteres.

Et XMLdokument skal som minimum være ”Welformed”, hvilket betyder at det overholder reglerne for opbygningen af XMLopmærkninger (f.eks. skal en begyndelsesopmærkning altid afsluttes med en afslutningsopmærkning).
Et XMLdokument kan desuden være ”Valid” (og dermed i sagens natur ”Welformed”) hvis der til dokumentet knytter sig en DTD (Data Type Definition) eller et ”Scheme” som definerer i hvilken rækkefølge disse opmærkninger kan/skal forekomme.
Hvis der hertil knyttes et Stylesheet som, via reference til opmærkningerne, definerer skrifttype, fontstørrelse mv. , så vil XMLdokumentet desuden kunne præsenteres på et utal af måder uden at ændre i selve XMLfilens dataindhold.

En DTD og et Stylesheet kan enten være integreret i XMLdokumentet (en fil) eller ligge som separate filer, som dermed kan benyttes af andre XMLdokumenter (nemt at opdatere alle dokumenter på en gang).

Man kan sige at XML intet kan i sig selv, men at man kan ALT med XML.
I tilknytning til XML er der efterhånden skabt et utal af ekstensions som understøtter og supplerer XML. Udover de generelle ekstensions som f.eks. Stylesheets, findes der mere specielle ekstensions f.eks XQL som giver mulighed for at definere SQL statements (forespørgsler i databaser) og MathXML hvormed det er muligt at angive matematiske udtryk og værdier og XHTML hvormed det er muligt at definere HTML version 4.0.
Fælles for disse ekstensions er at de alle er defineret via XML og altså ikke bryder med standarden.

XML har taget udgangspunkt i den grafiske verden (SGML), men er i en vis forstand ved at bevæge sig i andre retninger end oprindeligt tænkt, fordi XML er veldefineret, platformuafhængigt, maskinuafhængigt, smidigt og (meget snart!) meget udbredt .
XML ved at slå igennem specielt som meddelelsesformat mellem applikationer og mellem diverse typer af hardware (PC'er, WAP mv.). De alm. kendte transportprotokoller så som TCP/IP og HTTP vil fortsat blive benyttet, men meddelelsesindholdet vil blive XML og typen af modtagerens operativsystem og applikations-type vil være underordnet.

Som eksempel på anvendelsen af XML kan nævnes nogle af verdens største nyhedstjenester, som arbejder med udvikling af deres egne ekstensions til XML. Disse tænkes brugt således, at en given nyhed, XML opmærkes og lagres på en central server, hvorefter de forskellige medier via denne opmærkning vil kunne trække de relevante informationer ud i en form som passer til et ønsket medie. Det vil hermed være muligt for f.eks. en WAP telefon at trække på samme data som en nyhedsside på Internettet og en skreven side i en avis. Disse XMLekstensions fokuserer specielt på tekst, billeder, video, lyd og tabeller og kunne med tiden være interessante for Statens Arkiver fordi det kommer til at håndtere stort set samme filformater som et ESDHsystem.

Fremtiden

Hvis jeg skal give et bud på hvilken retning udviklingen indenfor filformater tager, så ville et forsigtigt bud på fremtidens filformater være; at kategorien billeder, lyd og video håndteres af filformater af MPEG

typen, og at stort set alt andet håndteres via XML.