PDF/A

Det at PDF/A formatet er definert som arkivformat for dokumenter i Norge skal være kjent for de fleste. Hva et PDF/A dokument egentlig er kan fremdeles være ukjent for mange. Artikkelen har som hensikt å fungere som en kort introduksjon til PDF/A formatet som arkivformat, og tar for seg PDF/A formatets formål, egenskaper, og ulikheter i forhold til PDF 1.4 standarden.

Av petter.pedryc@ika-trondelag.no

Ill: ePublicist/CreativeCommons: Attribution-NoDerivs 2.0 Generic.

Prinsippet bak PDF/A standarden er enkelt; bevaring. I dette ligger behovet for korrekt og lik fremvisning av dokumentet over tid. For å realisere dette behovet må man sette krav til hvor avansert funksjonalitet et PDF-dokument kan inneholde. Det som gjør PDF 1.4 standarden dårlig egnet for langtidsbevaring er at dokumentet på mange måter kan sies å fungere som en container for andre format, eksempelvis lyd og video. Avansert funksjonalitet i et dokument kan sies å sette grenser for hvor lenge dokumentet med sikkerhet lar seg lese, og motsatt. PDF/A dokumentet er med andre ord en kraftig forenkling av PDF 1.4 standarden, og setter følgende restriksjoner hva innhold angår:

- Audio and video content are forbidden.
- JavaScript and executable file launches are forbidden.
- All fonts must be embedded and also must be legally embeddable for unlimited, universal rendering. This also applies to the so-called PostScript standard fonts such as Times or Helvetica.
- Colorspaces specified in a device-independent manner.
- Encryption is forbidden.
- Use of standards-based metadata is mandated.
- External content references are forbidden.
- LZW and JPEG2000 image compressions are forbidden in PDF/A-1, but JPEG 2000 compression is allowed in PDF/A-2.
- Transparent objects and layers (Optional Content Groups) are forbidden in PDF/A-1, but they are supported in PDF/A-2.
- Provisions for digital signatures in accordance with the PAdES (PDF Advanced Electronic Signatures) standard are supported in PDF/A-2.
- Embedded files are forbidden in PDF/A-1, but PDF/A-2 offers the possibility to embed PDF/A files, allowing archiving of sets of documents in a single file.

Kilde: Wikipedia: PDF/A (http://en.wikipedia.org/wiki/PDF/A#Description)

Samtlige av disse kravene skal gjøre fremtidig rendering, eller gjengivelse på godt norsk, enklere. Tommelfingerregelen er at dokumentet ikke skal inneholde andre formater som krever ytterligere dokumentasjon for gjengivelse, samt at all informasjon nødvendig for å gjengi dokumentet enten befinner seg i PDF/A dokumentet eller i PDF/A standarden. Som en følge av dette er det naturlig at eksempelvis lyd og video ikke er tillat i PDF/A dokumentet, slik det er i PDF 1.4 dokumenter. Videre skal eksempelvis samtlige fonter (altså skrift typer) ligge i dokumentet, ikke i leseren, slik tilfellet er med både PDF 1.4 og Microsoft Word.

PDF/A dokumentet finnes i flere versjoner, med ulike egenskaper innenfor hver versjon. PDF/A-1 er den som er godkjent arkivformat i Norge. Det finnes også en PDF/A-2 og en PDF/A-3-versjon som skal være klar i løpet av høsten. Det er likevel ikke slik at den ene versjonen er bedre enn eller en forbedring av den forrige. Hver versjon er utviklet for å tilfredsstille litt ulike behov, men samtlige har som mål å fungere som arkivformat. Det finnes også en del andre spesielt tilpassete PDF-formater, samlebetegnelsen på disse er PDF/X. Eksempelvis er PDF/E spesielt utviklet for Engineering, altså tekniske tegninger.

Innenfor PDF/A-1 foreligger det to ulike format, såkalt A og B – en grei tommelfinger regel er at B står for «basic». Det er B som er den enkleste formen av disse to, den fokuserer på og garanterer kun for korrekt visuell representasjon. A-versjonen på sin side har i seg samtlige krav som B, men skal i tillegg evne å håndtere kontekst og mening. A håndterer eksempelvis det som kalles «typeset» på engelsk, altså inndeling av tekst og innhold i for eksempel paragrafer, avsnitt, den inneholder en egen metadata tag for dokumentets tittel etc. I praksis betyr dette at om man ønsker å gjenbruke teksten om 30 år, så kan man kopiere teksten rett ut av et PDF/A-1A dokument, og også få med formateringen. A-versjonen inkluderer også ting som tagging av bilder, altså man kan legge informasjon inn i bilder, slik at bildets opprinnelige innhold også er forståelig i fremtiden. Problemet med 1A versjonen er dog at konvertering til denne versjonen vanskelig lar seg automatisere, fordi dokumentet krever en del informasjon fra brukeren eller systemet før konvertering. 1B på sin side inneholder ingen «metadata» som krever spesiell brukerbidrag, her er det kun utseende som blir konvertert. Min antakelse er at de fleste norske Noark-5 systemene gjør seg nytte av 1B-versjonen. I de tilfeller kommuner ønsker å etterprøve dette så er det viktig at man forsøker å validere dokumentene i henhold til denne versjonen, ikke 1A.

PDF/A-2 utvider PDF/A-1, og tilbyr det som på engelsk heter transparency (gjennomsiktighet), layers (flere lag, dette hører gjerne sammen med gjennomsiktighet i dokumenter), samt JPEG2000-kompresjon for bilder. I forhold til norske behov er det nok JPEG2000-komprimering av bilder som er interessant. Dette er et tapsfritt kompresjonsformat som klarer å levere god kvalitet på til dels sterk komprimering. I tillegg så åpner dette formatet opp for at PDF/A-filer inneholder andre PDF/A-filer i seg. Akkurat denne muligheten kunne det vært aktuelt å benytte seg av i forhold til store saksdokumenter med mange vedlegg. I den grad jeg er kjent med PDF/A-2, kjenner jeg ikke til noen ulemper eller begrensninger som bryter med norsk bevaringsstrategi for elektronisk materiale. Dermed vurderer jeg det slik at denne versjonen kan bli et aktuelt arkivformat neste gang forskriftene revideres.

PDF/A-3 skal være ferdigstilt til høsten. Hovedpoenget her er at PDF/A-filen kan inneholde andre filer som ikke nødvendigvis behøver å være PDF/A-filer. Her kan man se for seg et PDF/A-dokument som fungerer som container for musikk, video og andre multimedieformater. Min mistanke er at

dette er en PDF/A versjon som har blitt utviklet med e-post i tankene. Bevaring av store mengder e-post har vært en stor utfordring i tilfeller hvor arbeidet gjøres i etterkant og ikke fortløpende, særlig på grunn av alle vedlegg som ofte følger med. Til tross for at denne versjonen har blitt lagt under PDF/A-paraplyen, er det nok fremvisning og ikke bevaring som har stått i fokus her. PDF/A-3 presenterer en enkel løsning for å gjøre e-post tilgjengelige for en kort periode, men såfremt bevaring av samtlige vedlegg som følger med inne i dokumentet ikke adresseres så vil disse dokumentene få en langt kortere utløpsdato enn versjon 1 og 2. På grunnlag av nettopp dette ser jeg ikke for meg at formatet vil bli tillat som arkivformat i Norge. For privat næringsliv som ofte har 10 års bevaringsplikt på blant annet økonomi vil nok dette formatet være svært attraktivt.

For samtlige Noark-5 systemer i dag gjelder likevel at disse skal evne konvertering til PDF/A-1, som pr. 15.5.2012 er gjeldende arkivformat for dokumenter. Kommuner bør avklare hvilken PDF/A-1-versjon systemet konverterer til, og sørge for jevnlige tester hvor dokumenter valideres i henhold til aktuell standard. IKA Trøndelag vil ikke anbefale applikasjoner for validering, men dersom man går til PDF Association sine hjemmesider så vil man finne både veiledere samt anbefalinger til programvare. Dersom man ønsker å fordype seg videre i PDF/A er PDF Association sin artikkel ved navn «PDFA in a Nutshell» en god plass å begynne.

- PDF in a Nutshell av PDF Association (ekstern lenke)
- Validering av PDF/A av PDF Association (ekstern lenke)

Til slutt må jeg få lov til å si at artikkelen er et resultat av min forståelse av og min kunnskap om PDF/A. Dersom noe skulle vise seg å være feilaktig eller manglende tar jeg gjerne imot tilbakemeldinger på det. Artikkelen vil i så tilfelle revideres i henhold til tilbakemeldingene.