Dokumenttimuotojen migraatio – alustus

KDK-pitkäaikaissäilytys 2013 -seminaari
6.5.2013 / Juha Lehtonen

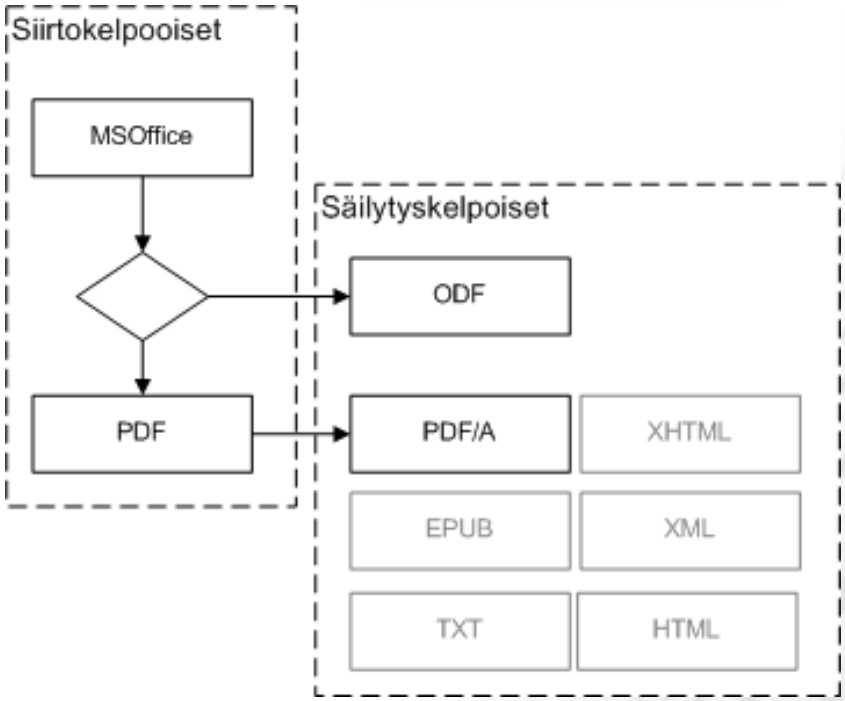
Tiedostomuodot ja PAS

- KDK:ssa on olemassa määritys säilytyskelpoisille ja siirtokelpoisille tiedostomuodoille
  - Säilytyskelpoiset muodot otetaan PAS-ratkaisuun vastaan ja säilytetään sellaisenaan
  - Siirtokelpoiset muodot otetaan PAS-ratkaisuun vastaan, mutta ne migroidaan säilytyskelpoisiksi ennen säilytyksen aloittamista
  - Muita muotoja ei oteta vastaan

Migraatioiden tilanne

- PAS-ratkaisun toteuttamissunnitelman mukaan vuonna 2013 PAS-ratkaisussa aloitetaan bittien säilyttäminen
  - Ei sisällä migraatioita
- Vuosina 2014-2016 toteutetaan ymmärrettävyyden säilyttämisvaihe, joka sisältää migraation
- Nyt ollaan siis migraatioiden suhteen valmisteluvaiheessa
  - Vaikka fokus on seuraavan muutaman vuoden ajassa, niin migraatioista täytyy keskustella jo nyt

Dokumenttimuotojen migraatio



Dokumenttimuodot

- MS Office 97-2003 -muodot
  - Binääridataa, ei luettavissa ASCII-tekstinä
  - Tiedostopäätteet: .doc, .xls ja .ppt
- MS Office 2007-2013 -muodot
  - XML-pohjainen: Office Open XML (OOXML)
  - ZIP-pakattu
  - Tiedostopäätteet: .docx, .xlsx ja .pptx
- OpenDocument-muoto (ODF)
  - XML-pohjainen
  - ZIP-pakattu
  - ISO/IEC-standardoitu (2006)
  - Tiedostopäätteet: .odt, .ods ja .odp

MS Office 2007 –muotojen ja OpenOffice-muotojen eroja

käytettäessä MS Office 2007:ää

Lähde

- Perustuu Microsoftin tekemiin taulukoihin ODF-formaateista*)
  - Tukeutuu MSOffice 2007:n ODF-käsittelyyn.
    - Taulukot esittelevät nimenomaan MSOfficen käsittelytapoja, ja voivat olla jo osin vanhentuneita.
  - Taulukot sisältävät yhteensä yli 300 kohtaa, silti ei ole oikeastaan miltään osin yksityiskohtainen
  - Lähdettä lukiessa ei ole aina selvää, johtuvatko luettelon asiat
    - ODF-formaatin omista ominaisuuksista
    - MSOfficen ODF-tallennuksen ominaisuuksista
    - MSOfficen ODF-avauksen ominaisuuksista
  - Joka tapauksessa: Näiden asioiden kanssa saattaa olla ainakin joitakin ongelmia

*) Lähde: http://office.microsoft.com/

Tekstimuotoilut

- Eri fontteja ja kokoja, sekä kursivointia, lihavointia, alleviivausta ja yliviivausta tuetaan.
- Word:
  - Seuraavia tekstin suuntamuotoja tuetaan:
    - kirjoitus vasemmalta oikealle, rivit ylhäältä alas
    - kirjoitus oikealta vasemmalle, rivit ylhäältä alas
    - kirjoitus ylhäältä alas, rivit oikealta vasemmalle.
  - Taulukon solussa käytettävää tekstin suuntaa ei tueta.
  - Kaikkia rivityylejä ja rivin pään tyylejä ei tueta ODF-muodossa. Tyylit, joita ei tueta, tallentuvat oletustyylinä.
  - Korostus muunnetaan merkin taustaväriksi.
  - Ala- ja loppuviitteet: Mukautettuja erotinmerkkejä ei tueta.

Tekstimuotoilut

- PowerPoint:
  - Tekstiruutujen fonttikoko, joissa on Sovita-asetus käytössä, määritetään yhteen kokoon.
  - Ei-tuettuja kansainvälisen tekstimuotoilun ominaisuuksia:
    - Pystysuoran tekstin (270, pinottu) kierto
    - Tietyt Itä-aasialaiset rivitykset
    - Jotkin kansainväliset numerointijärjestelmät (yhdistetään luettelomerkeiksi tai länsimaiseen numerointiin)
    - Jaettu tasaus (rivit täytetään kaikkien merkkien väleistä sanavälien asemasta).

Reunat ja viivat

- Reunaviivat ovat tuettuja, mutta ne eivät ole välttämättä samannäköisiä.
- Kaikkia reunatyylejä ei tueta. Reunatyylit, joita ei tueta, tallentuvat oletusreunatyylinä (musta, yhtenäinen viiva).
- ODF ei tue kaikkia viivan ja viivanpään tyylejä. Tyylit, joita ei tueta, tallennetaan oletusarvon mukaisesti mustana, yhtenäisenä viivana ja avoimena nuolena.
- Sävytyksen kuvioita ei tueta. Kuvan reunaviivatyylejä ei tueta, ne muunnetaan yhtenäiseksi viivaksi.

Värimuutokset

- Teemojen värit muunnetaan vastaavaan ODF-ominaisuuteen.
- Jollakin kuvan värien muuttamisella ei ole vastaavaa ominaisuutta, joten se tasoitetaan. Kuva näyttää samalta, mutta väriä ei voi enää muuttaa eikä värin muutosta poistaa.
- Täyttöominaisuudet, joita ei tueta:
  - kuvatäyttö, liukuväritäyttö, kuviotäyttö tai tyhjän tekstin täyttö
  - jotkin tekstin ulkovarjostustehosteet; teksti, jossa on kuva, liukuväri, kuvio tai tyhjät täytön ääriviivat
  - tekstin ääriviivan väri ei voi olla eri kuin täytön väri
- PowerPoint: Dian taustan täyttöä ei tueta. Jos liukuväritäytössä on enemmän kuin kaksi rajaa, ensimmäisen kahden rajan jälkeen tulevat muut rajat poistuvat.

Asettelu (Word)

- Tekstisarakkeet: Jotkin osaan liittyvät ominaisuudet, kuten ylä- ja alareunukset, ylä- ja alatunnisteet, reunat tai rivinumerointi, saatetaan menettää.
- Taustan muotoilu, teemat, vesileima tai asiakirjan rakenneosat eivät ole tuettuja.
- Joidenkin muokkausruutujen, kehysten ja muotojen sijainti voi muuttua käytetyn ankkurin tyypin mukaan.

Asettelu (PowerPoint)

- Teeman tiedot menetetään ja sijoitetaan perustyyleihin/asetteluihin.
- Joitain visuaalisia eroja voi nähdä tekstin tasauksessa, ankkuroinnissa tai rivityksessä muissa ODF-sovelluksissa.
- Ylivuoto saattaa näkyä eri tavoin eri ODF-sovelluksissa.
- Asetteluja käsitellään perustyylien tavoin, kun ne avataan muissa ODF-sovelluksissa.
- Paikkamerkit tallennetaan ja säilytetään ODP-tiedostoissa. Kaikki paikkamerkit jatkavat muotoiluominaisuuksien perimistä asettelujen ja dian perustyylien vastaavista paikkamerkeistä. Kaikki muotoilun ohitukset säilytetään myös.

Kehykset ja kentät

- Kehykset muunnetaan tekstikehyksiksi. Joitakin reunuksen alueiden ankkureita ei tueta. Sisältö, jota ei tueta, aiheuttaa kehysten menettämisen, mutta itse sisältöä ei menetetä. Esimerkiksi seuraavaa sisältöä ei tueta: taulukot, automaattiset muodot, tekstikehykset, kehykset ja SmartArt-grafiikka.
- Kentät, joita ei tueta, muunnetaan normaaliksi tekstiksi (esim. SEQ-kentät).
- Tiedostoissa, jossa on ylä- tai alatunnistekenttiä, mukaan lukien päivämäärä/aika- ja sivunumerokentät, tunnisteet muuttuvat tekstiruuduiksi, joita ei enää yhdistetä ODF-sovelluksen ylä-/alatunnisteominaisuuksiin.
- Jotkin jatkuvien osan vaihtojen ominaisuudet, kuten ylä- tai alareunukset, ylä- tai alatunnisteet, reunat ja rivinumerointi, voidaan menettää.
- Ristiviitteet ja sivunumerointi on tuettu.

Luettelot

- PowerPoint: Muut ODF-sovellukset saattavat muuttaa numerointia tai luettelomerkkejä tai eivät tue sitä ollenkaan.
- Word: Numeron tai luettelomerkin ja tekstin väli voi olla hieman erilainen. Luettelokohteiden välejä suurennetaan asiakirjan riviväliä vastaavaksi.
- SEQ-kentät muunnetaan normaaliksi tekstiksi:
  - Jos käyttäjä lisää uusia kuvaotsikoita, otsikon numero ei suurene automaattisesti.
  - Kuvaotsikkoluettelo ei ole tuettu ominaisuus.
  - Sisällysluettelon kohteet, jotka on otsikoitu SEQ-kentän avulla, menetetään.
  - Lähdeluettelo-osa muunnetaan normaaliksi tekstiksi.

Taulukot

- Excel:
  - Taulukkotyylejä ei tueta
  - summarivejä ei tueta
  - joitakin Pivot-taulukon asetteluja, kuten kutistettua akselia, ei tueta.
- Word:
  - Taulukoita, joissa on enemmän kuin 64 saraketta, ei tueta.
  - Teemojen muotoilu muunnetaan solutason muotoiluksi.
- PowerPoint:
  - Taulukot muuttuvat kuviksi eikä niitä voi muokata.

Kaaviot (Excel)

- Kaaviossa olevat objektit muunnetaan kaavion ja objektit sisältäväksi ryhmäksi.
- SmartArt-kaaviot muunnetaan muotoryhmäksi.
- Ei tueta: Arvopisteen otsikoiden viiteviivat, arvotaulukot, arvoviivat, kaavioissa olevat muodot, kaaviotaulukot, pivot-kaaviot, täytetyt säteittäiset kaaviot, ylimpien ja alimpien arvojen rivit, ympyrä- tai palkkikaavio sektorista

Muodot

- PowerPoint: Muodon piilottaminen on tuettu.
- Muodossa käytetty hyperlinkki ei ole tuettu
- 3D-muotoja tuetaan Wordissa
  - Ei tueta Excelissä tai PowerPointissa.
- 3D-kuvien asetukset eivät säily

Objektit

- ODF tukee PowerPointin, mutta ei Wordin tai Excelin, näkymättömiä objekteja. Myös dian voi piilottaa.
- Ei tueta: Excel-lomakkeiden ohjausobjektit, Kameratyökalu tai liittäminen kuvalinkkiobjektina, Kirjautumisriviobjekti
- Kaavat (Kaavaeditori) on tuettu ominaisuus
- WordArtia tuetaan Wordissa; mutta ODF ei tue WordArt-asetuksia Excelissä tai PowerPointissa.
  - Muunnetaan tallennettaessa muokkausruuduksi. Tekstin perusväri säilyy, mutta WordArt-tehosteet katoavat.
- PowerPointissa tuettuna:
  - Media (elokuvat/äänet), Upottaminen: WAV-tiedostot

Animaatiot (PowerPoint)

- Mukautetut esitykset, selostus, ajastaminen tuettu
- Animaatioiden perusominaisuudet tuettu
  - Ajoitus ja viiveet, käynnistettävät animaatiot, media-animaatiot, tekstianimaatiot
- Aloitus- ja lopetusanimaatiot muutetaan ODF:n ilmestymis- ja katoamisanimaatioiksi
- Ei tueta: animaatioiden ääniä, dian perustyyliin tai asetteluun liittyviä animaatioita, OLE-toimintojen animaatioita, skaalautuvia animaatioita, SmartArt-animaatioita, värin muuttamiseen liittyviä animaatioita
- Joitakin siirtymiä ei tueta

Sekalaista 1/2

- Hyperlinkit:
  - Hyperlinkkityylejä ei käytetä ODF:ssä (toimivat silti)
  - Tavallisia hyperlinkkejä tuetaan, mutta osoittamalla valittavia (hover) hyperlinkkejä ei tueta.
- Ohjelmointi:
  - ActiveX-komponentteja ja makroja ei tueta.
  - VBA-makroja tuetaan.
- Suojaus:
  - Salattuja tiedostoja ei voida tallentaa ODF-muotoon
  - Yhteiskäytön sisältöoikeuksien hallintaa ei tueta
  - Excel-taulukon suojaus ilman salasanaa toimii, salasanasuojattua taulukkoa ei voida tallentaa ODF-muotoon

Sekalaista 2/2

- Kaikki päivämäärätyypit on sisällytetty, mutta muut ODF-sovellukset saattavat muuttaa ne oletustyypiksi.
- Wordin erilliset XML-ominaisuudet eivät ole tuettuja.
- Kommentteja ei tueta
  - Excelissä solun kommentin sisältö säilyy, mutta ei välttämättä muotoilu.
- PowerPoint:
  - Muistiinpanojen ja tiivistelmän perustyylit sekä itse muistiinpanot tuettu
  - Sivuasetukset eivät välttämättä ole täsmällisiä.

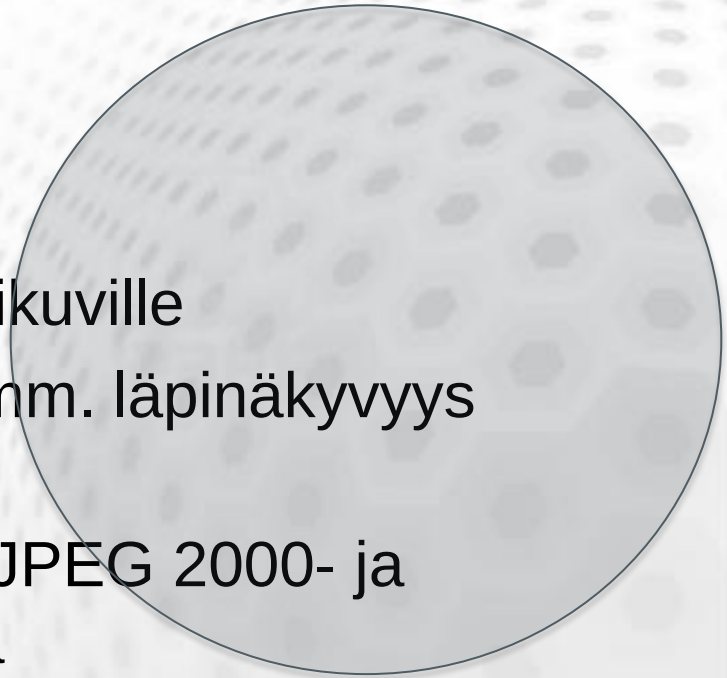
PDF ja PDF/A

Tiedostomuoto PDF

- PDF: Portable Document Format
- Kehittäjä: Adobe
- Eri versioita: 1.3-1.7, kehitteillä 2.0
- ISO-standardiksi 2008 (versio 1.7)
- Tiedostomuotossa dokumentti rakentuu eri typpisistä objekteista
  - Objekteihin voidaan asettaa läpinäkyvyys-ominaisuuksia
- Mukana XML-pohjainen metadatakokonaisuus
- Muotoon voi upottaa tiedostoja

Tiedostomuoto PDF

- 14 perusfonttia (mutta lisää voi upottaa)
- Kryptaus, sähköinen allekirjoitus
- Interaktiiviset elementit
  - Esim. lomakkeet
- Kuvat
  - Tuki sekä rasteri- että vektorikuville
  - Kuvadatan ominaisuuksina mm. läpinäkyvyys ja väriavaruus
  - Rasterikuvissa mm. JPEG-, JPEG 2000- ja LZW-pakkaukset mahdollisia
  - Näistä lisää huomenna



Osittain läpinäkyvä ympyrä

Officesta PDF:ksi

- Tästä on monilla paljon kokemuksia
- Runsas kysyntä on johtanut myös hyvään muunnostyökalujen kehitykseen
  - Ja yleensä muunnos siis onnistuu aika hyvin
- Seuraukset samantapaiset kuin tulostettaessa paperille
  - Ulkoasu säilyy
  - Muokattavuus häviää
  - Jotkut objektit tai osiot voivat vuotaa usealle sivulle
    - Erityisesti taulukkolaskentadokumentit
  - Interaktiiviset osiot eivät välttämättä toimi
  - Fontit voivat tietyissä tilanteissa muuttua

PDF/A

- Erityisesti säilytystä ajatellen tehty muoto
- ISO-standardoitu
- Riisuttu versio PDF:stä
  - Poistettu kaikki PDF:n ominaisuudet, jotka potentiaalisesti tekevät säilyttämisestä mahdotonta
- Eri versioita:
  - PDF/A-1a, PDF/A-1b (2005)
    - Perustuu PDF 1.4:een
  - PDF/A-2a, PDF/A-2b ja PDF/A-2u (2011)
    - Perustuu PDF 1.7:ään
  - PDF/A-3 (2012)

PDF/A-1b-muodon rajoitukset PDF-muotoon nähden

- Audio- ja videosisältö on kielletty
- JavaScript-sovellukset tai ohjelmien käynnistämiskomennot ovat kiellettyjä
- Kaikki fontit pitää olla sisällytettynä
  - Koskee myös PDF:n perusfontteja
  - Laillisuusnäkökulma huomioitava
- Väriavaruudet on ilmaistava standardeina (ei laiteriippuvina) muotoina
  - Tästä huomenna lisää
- Kryptaus (salaus) on kielletty

PDF/A-1b-muodon rajoitukset

- Metadatatiedot ovat pakollisia
- Viittaukset ulkopuoliseen sisältöön ovat kiellettyjä
- LZW- ja JPEG2000-kompressiot kiellettyjä
- Läpinäkyvyysominaisuudet kiellettyjä
- Upotetut tiedostot kiellettyjä

PDF/A-1a-muodon rajoitukset

- **Kaikki PDF/A-1b-muodon rajoitukset**
- Dokumentin hierarkia pitää olla sisällytetty
- ”Tagged PDF” pakollinen
  - Esim. kuvilla pitää olla mukana selite (joka näytetään, jos kuvaa ei pystytä esittämään)
  - Symboleilla pitää olla vaihtoehtoinen teksti
- Unicode-merkistö pakollinen
- Kielimääritykset pakollisia

PDF/A-2

- PDF/A-2 antaa lisää vapauksia
  - JPEG 2000 kuvat sallittuja
  - Läpinäkyvyys objekteissa sallittu
  - Sähköiset allekirjoitukset mahdollisia
  - Voidaan sisällyttää useita PDF/A-2 dokumentteja samaan tiedostoon
- PDF/A-1-dokumentti on myös PDF/A-2-kelpoinen
  - PDF/A-2a ja 2b muotojen erot kulkevat käsi kädessä PDF/A-1a ja 1b muotojen kanssa
  - PDF/A-2u on sama kuin PDF/A-2b, mutta unicode-merkistövaatimuksen kanssa

PDF/A-3

- Ei KDK:ssa säilytys-/siirtokelpoinen
- On täysin sama kuin PDF/A-2, mutta antaa mahdollisuuden sisällyttää dokumenttiin mitä tahansa tiedostomuotoja

Officesta PDF/A:ksi

- Suora muunnos ei välttämättä toimi kovin hyvin
  - Esimerkiksi, jos objekti on osittain läpinäkyvä, muunnos saattaa poistaa objektin kokonaan
- Ainakin toistaiseksi on parempi muuttaa Office-muodosta PDF:ksi ja sen jälkeen PDF/A:ksi
  - Esimerkiksi, läpinäkyvyys on PDF:ssä parametri, jonka muunnos usein vain poistaa. Objekti jää näkyviin ilman läpinäkyvyyttä