docx-formaatti

Tiedostot

Microsoftin koulutus aiheeseen: http://office.microsoft.com/en-us/training/open-xml-ii-editing-documents-in-the-xml-RZ010357030.aspx

* [Content_Types].xml - sisältää listauksen kaikista dokumentin xml-tiedostoista
* tänne tulee kaikki tiedostoformaatit, joita on docx-tiedostossa käytetty, esim. png-kuvat
* word-kansio
* _rels-kansio
* document.xml.rels - sisältää tietoja tiedostojen välisistä suhteista ja esimerkiksi kuvista. Yksilöi tiedostot rId-arvolla.
* hyperlinkit tulevat tänne erillisinä viitteinä
* document.xml - itse dokumentti ja sen sisältö
* styles.xml - sisältää tiedot tyyleistä
* fontTable.xml - sisältää tietoa dokumentissa käytetyistä fonteista (!)
* voi sisältää myös muita tiedostoja, riippuen siitä mitä kaikkea dokumentti sisältää
* esimerkiksi numbering.xml, header1.xml, footer1.xml...
* theme-kansio
* theme1.xml - sisältää tiedot oletusfonteista? (major-fontti & minor-fontti = otsikkofontti & tekstifontti?)

* docProps - kansio
* app.xml - sisältää tietoa mm. sanamäärästä, ohjelmaversiosta, muokkauskestosta (integer)
* core.xml - sisältää tiedot tekijästä, viime muokkaajasta, luomisajasta, viime muokkausajasta

Sisältö

document.xml - rakenne (sisältää myös esimerkkejä):

* <w:document>
* <w:body> voi sisältää useita kappaleita
* <w:p> - paragraph voi sisältää useita run-elementtejä
* <w:pPr> paragraph properties (vaikuttaa koko kappaleeseen)
* <w:rPr> run properties (vaikuttaa kaikkiin runeihin)
* <w:r> - run
* <w:rPr> run properties (vaikuttaa tähän runiin)
* <w:t> - text range - teksti
* <w:lastRenderedPageBreak/> - Position of Last Calculated Page Break
<w:t>SISÄLLYSLUETTELO</w:t>

* document sisältää yhden bodyn
* body voi sisältää useita kappaleita

* kappale voi sisältää useita "ajoja" (run)
* ajon sisältö voi olla muutakin kuin tekstiä

* xml:space="preserve" - säilyttää välilyöntimerkit
* xml:space = "preserve": "since leading and trailing whitespace is not normally significant in XML; some runs require a designating specifying that their whitespace is significant via the xml:space element."

http://stackoverflow.com/questions/116139/how-can-i-read-a-word-2007-docx-file

Riippuvuudet

Dokumentin osien väliset riippuvuudet ilmoitettu rels-tiedostoissa. Esimerkiksi: dokumentissa on kuva - xml:ssä kuvan paikalla on elementti, jossa rid-kenttä - rels-tiedostossa vastaava id viittaa media-kansiossa olevaan kuvaan.

Kuvista lisää täällä: http://office.microsoft.com/en-us/training/replace-document-images-directly-in-the-xml-RZ010357030.aspx?section=14

Tyylit

Lähestulkoon kaikki tyylit peritään Normal-tyylistä, joka perii oletusasetukset. Leipäteksti on xml:ssä "Leipteksti", engl. "Body Text".

Tyylit määritellään styles.xml-tiedostossa, ja niihin viitataan document.xml-tiedostossa. Tekstin suoran muotoilun seurauksena ei viitata tyyliin, vaan tyylimääritykset tulevat suoraan document.xml-tiedostoon.

Oletusarvot styles.xml-tiedostossa:
* <w:docDefaults>
* Oletusfontti, oletuskieli

Leski- ja orporivien esto oletuksena päällä. Jos sen ottaa pois päältä, löytyy asetus dokumentista.
* <w:widow... val=0>
Tekstin keskitys on oletuksena vasemmalla. Jos sitä muuttaa, asetus löytyy dokumentista.
* <w:jc w:val=" " />
Riviväli on oletuksena yksi. Jos sitä muuttaa, asetus löytyy dokumentista.
* <w:spacing w:line=""

Erilaisia tyylityyppejä:
* Paragraph styles - kappaleen tyyli määrittelee esim kappaleen ryhmittymisen

* Character styles - tekstin tyyli määrittelee esim tekstin fontin, boldin, värin jne
* Linked styles (paragraph + character) - linkitetty tyyli: kappaletyyliin on linkitetty tekstin tyyli
* Table styles
* Numbering styles
* Default paragraph + character properties

* Latent styles: tyylejä dokumentin templatessa, joita ei ole vielä käytetty kyseisessä dokumentissa

Paragraph styles (kappaletyylit):
* Voi määritellä sekä kappaleen ominaisuuksia (w:pPr, paragraph properties) ja tekstin ominaisuuksia (w:rPr, run properties)
* kappaletyyliin voi viitata vain kappale-elementistä (<w:p>). Jos tyyli määrittelee myös tekstin ominaisuuksia, kappaleen ajot perivät ominaisuudet.

Character styles (tekstityylit):
* Määrittelee ajon (run) ominaisuuksia
* Voi viitata vain run-elementistä (<w:r>).

Linked styles:
* Esimerkiksi kappaletyyli, joka voi normaalisti sisältää sekä kappaleen että tekstin tyylitystä
* Lisäksi linkitys (<w:link w:val="TestLinkedStyleChar"/>) erilliseen tekstityyliin (character style)
* Samaa tyyliä voi käyttää sekä kappaleelle, jolloin käytetään kappaletyylin määrityksiä, sekä yksittäiselle run-elementille, jolloin käytetään tekstityylin määrityksiä

Default paragraph + character properties
* "Although this is not a style in the strict sense of the word (because this property set cannot
directly be applied to text) it defines the basic set of formatting properties which are inherited by paragraphs and runs in the document."
* Kts. tyylien perintä

Tyylien perintä:
* Jokainen tyylimääritys voi periä jonkun aiemman samantyyppisen tyylimäärityksen (esim. kappaletyyli voi periytyä jostain toisesta kappaletyylistä, vrt. Text Body -> Bold Text Body)
* Perijä voi ylikirjoittaa perityn elementin tyylimäärityksiä
* basedOn-elementti
* Koko tyylimäärityksen selvittämikseksi täytyy käydä läpi tyylipuuta kunnes ei enää löydy basedOn-elementtiä
* Oletusasetukset paikallisia ja ohjelmakohtaisia? Onko ongelma?

Fontit

* Suora asettaminen: <w:rFonts w:ascii="Arial Black" w:hAnsi="Arial Black" w:cs="Arial" w:eastAsia="SimSun"/>
* ascii, high ANSI, complex script, eastAsia

* Oletukset (Normal-tyylin):
* styles.xml <docDefaults>

* Teemafontit:
<w:rPr>
<w:rFonts w:asciiTheme="minorHAnsi" w:hAnsiTheme="minorHAnsi" />
</w:rPr>
* viittaus theme-kansion themeX.xml-tiedostossa määritettyyn fonttiin

* font table - sisältää tietoa dokumentissa käytetyistä fonteista
    * jos haluttua fonttia ei löydy, font tablesta ohjelmat löytävät korvaavia fontteja

Sivun asetukset (document.xml)

* <w:sectPr>-elementissä
* <body>-elementin viimeisenä lapsena
* muualla <w:p>-elementin sisällä, missä osa vaihtuu
* headerReference
* footerReference
* Sivun korkeus ja leveys
* Marginaalit

Note: http://office.microsoft.com/en-us/training/understanding-formatting-as-shown-in-document-xml-RZ010357030.aspx?section=6

1 inch = 1440 twips

Muut asetukset

* Automaattinen tavutus: settings.xml - <autoHyphenation />