

Microsoft Word 97:n käyttö HTML:n tuottamiseksi

Timo Hellgren

17. joulukuuta 1998

1 Johdanto

Microsoft Word 97 on tällä hetkellä yleisin Suomessa käytettävä tekstinkäsittelyohjelma. Syynä yleistymiseen on Office-ohjelmistopakettien käyttöön siirtyminen useimmissa yrityksissä ja virastoissa. Office sisältää lisäksi taulukkolaskennan *de facto* -standardin, Excelin, ja esitysten tekoon tarkoitetun suosituksen ohjelman, Power Pointin.

Word 97:llä voidaan tallettaa dokumentti HTML-tiedostoksi. Tähän ei tarvita mitään apuohjelmia, vaan toiminto on osa ohjelmaa.

2 Kokeiluaineisto

Kokeiltavana olivat seuraavat dokumentit:

1. Muisti-projektin loppuraportti. Tehty Word 97:llä. Mukana oli lisäksi grafiikkaa Corel Draw -formaattissa.

3 Dokumentin rakenne

Word 97:n HTML-konversio tuottaa tiedostosta vastaavan HTML-tiedoston. Lyhyissä dokumenteissa tämä on aivan riittävä toiminto, mutta pidemmissä töissä tuntuisi paremmalta jakaa alkuperäinen dokumentti useampaan HTML-tiedostoon, jotka olisi linkattu keskenään.

Muistin raportti on yli 100-sivuinen dokumentti, jonka HTML-dokumentista tuli niin pitkä, että se ei edes mahtunut muistiin Netscapen Composer-editorissa. Jouduin jakamaan tiedoston käsin niin, että jokainen luku oli

omassa tiedostossaan. Tämän jakamisen jouduin tekemään EMACS-editorissa, jonne HTML-tiedosto sentään mahtui.

Lisäksi kaikki navigaationuolet eri osien välillä oli tehtävä käsin. Tätä työtä voi tietenkin rationalisoida myös Netscapen editorissa hieman, mutta käytännössä koko dokumentti oli käytävä käsin läpi. (Tämän olisin joutunut tekemään joka tapauksessa muiden käännösvirheiden vuoksi.)

Kokeilin myös verkosta saatavia apuohjelmia, joilla Wordin dokumenttien HTML-muunnos pitäisi onnistua paremmin. Ne olivat kuitenkin käytännössä hitaita ja kömpelöitä VBA-makrokokoelmia, jotka kaikki eivät edes suostuneet toimimaan. Lisäksi kokeilin jonkin verran ohjelmaa HTMLTrans, jolla tiedosto voidaan jakaa osiin ja tyyleille voidaan määritellä vastaavat HTML-käskyt. Ohjelma on kuitenkin turhankin kallis ja kokeiluversion kokeiluajan umpeuduttua päädyin odottelemaan Wordin seuraavaa versiota, jonka beta-versio on piakkoin saatavilla myös Suomessa ja jonka HTML-kykyjen pitäisi ennakkotietojen mukaan olla parempia kuin nykyiset. Toivottavasti Office 2000 lisää mahdollisuuden jakaa dokumentti useampaan HTML-tiedostoon halutuista kohdista lisäten samalla automaattisesti halutunlaiset navigaatio-uolet.

4 Metatagit

4.1 Merkistö

Word 97 lisää dokumenttiin metatagin, joka kertoo käytetyn merkistön. Tämä on hyvä piirre niille, jotka surffaavat eri kielisillä sivuilla, sillä kyseinen metatagi näyttää sivun uusimmilla selaimilla automaattisesti oikealla merkistöllä. Tarvittavat kirjasimet (esim. kyrilliset kirjaimet) tulee olla kuitenkin asennettuna käyttöjärjestelmän tai selaimen tarjoamin keinoin.

Word 97:n merkistönä on kuitenkin ”Windows-1252”, joka on siis Windowsin länsieurooppalaisten kielten koodisivu. Tästä ei pitäisi syntyä ongelmaa, jos HTML-sivua luetaan ainoastaan Windowsissa, mutta se ei ole ISO:n standardin mukainen ratkaisu. Internetissä perusmerkistöksi on nimittäin sovittu ISO Latin-1 -merkistö, joka ei ole sama asia kuin Windowsin koodisivu 1252, vaikka Microsoft yrittääkin antaa sellaisen vaikutelman. ISO Latin-1 merkistöä pitää siis käyttää jos kirjoittaa englanniksi, suomeksi, ruotsiksi, saksaksi, ranskaksi, italiaksi, espanjaksi, hollanniksi, tanskaksi, norjaksi tai islanniksi.

Väärän merkistön ongelma poistuu, jos sivun avaa Netscapen Composer -editorissa. Se käyttää oletusmerkistön arvona iso-8859-1, joka on ISO Latin-1 merkistön virallisen standardin nimi. Composer lisää kyseisen arvon

vastaavaan tagiin. Toinen vaihtoehto on korjata asia käsin.

ISO:n merkistö on käytössä mm. UNIX:ssa, jossa ei teknisten rajoitusten vuoksi voida käyttää kaikkia Windowsin merkistön koodipaikkoja, esim. Windowsin ajatusviiva on merkki, jonka paikalla UNIX:ssa on jokin kontrollikoodi. Niinpä tekstistä joudutaan poistamaan käsin sellaiset merkit, jotka eivät kuulu kyseiseen ISO:n merkistöön, esim. ajatusviivaa ei voi käyttää, sillä se ei valitettavasti sisälly ISO Latin-1 -merkistöön.

Mikäli teksti on kirjoitettu jollain muulla merkistöllä, voidaan käyttää mahdollisesti Word 97:n tarjoamaa merkistöä suoraankin. Esimerkiksi venäjän osalta on tilanne sellainen, että Windowsin vastaava merkistö 1251 on hyvin yleisesti käytössä Internetissä. ISO:n vastaavaa merkistöä ei käytä ilmeisesti juuri kukaan, sillä UNIX:ssa kyrillisen merkistön standardina on KOI-8, joka on (omituinen) venäläinen standardi.

4.2 Luettelointitiedot

Mikäli Wordin dokumenttiin on tallennettu ”ominaisuudet”-toiminnolla tekijän ja dokumentin nimi sekä asiasanoja, siirtyvät nämä tiedot automaattisesti myös HTML-dokumenttiin vastaaviksi META-tageiksi. Jos halutaan kuitenkin käyttää Dublin Coren vastaavia META-tageja, on nämä tehtävä itse esimerkiksi HTML-editorissa.

5 Leipäteksti

HTML-dokumentissa kappaleiden ensimmäistä riviä ei sisennetä. Lisäksi kappaleet erotetaan tyhjällä rivillä toisistaan. Tämä ei aiheuta ongelmia HTML-konversiossa.

Lihavoitu ja kursiivi teksti korvataan vastaavilla HTML-tageilla. Samoin myös alleviivaus ja yliviivaus. Word näyttää tosin konversiossa sekoilevan jonkin verran, sillä toisinaan muotoilu jää päälle tai ei siirry lainkaan HTML-versioon.

Tabulaattorilla tehdyt sisennykset eivät toimi oikein HTML-sivuilla, joten kaikki taulukot ja sisennykset on tehtävä Word-dokumenttiin Wordin taulukkokäskyillä.

Kentistä ainoastaan näytöllä näkyvä tulos siirtyy HTML-dokumenttiin. Esimerkiksi automaattisesti päivittyvästä päivämäärästä HTML-dokumenttiin siirtyy se päivämäärä, jolloin konversio tehdään.

6 Otsikot

Word 97:n HTML-konversiossa otsikot näkyvät samanlaisina HTML-tiedostossa. Kirjoittajan ei tarvitse käyttää edes tyylejä, sillä HTML-konversiossa säilytetään muotoilukäskyt. Ongelmana on se, että HTML-tiedostossa ei käytetä HTML:n otsikkotageja, vaan otsikot tehdään pelkästään muotoilukäskyillä. Mikäli käytetään jotain muuta kirjasinta kuin ”Times New Roman” ei ole varmaa, että muotoilu säilyy lukijoiden selaimissa kuitenkaan samanlaisena.

Wordin automaattisesti numeroidut otsikot on numeroitu myös HTML-tiedostossa lisäämällä numero otsikon eteen. Tämä johtuu siitä, että HTML ei tue automaattisesti numeroituja otsikoita.

Muistin loppuraportissa kolmannen tason otsikoissa näytti olevan jotain vaikeuksia, sillä Word oli tulkinut ne HTML-tiedostoon jossain tapauksissa numeroidun luettelon kohdiksi. Virhe piti korjata käsin.

7 Hyperlinkit

Word 97:ssä on oma tyyli hyperlinkeille. Lisäksi ohjelma muuttaa tekstissä hyperlinkin näköisen osuuden automaattisesti kirjoitettaessa hyperlinkiksi. Toiminto ei kuitenkaan ole idioottivarma. Ohjelma erehtyy luulemaan mm. lyhennettä URL linkiksi silloin kun sitä seuraa kaksoispiste. Englannissahan tämä ei tuota ongelmia, mutta kylläkin suomessa.

Kun dokumentti muutetaan HTML-tiedostoksi on tarkistettava ovatko kaikki linkit järjellisiä. Turhat linkit täytyy poistaa yksinkertaisesti käsin HTML-editorissa.

Word 97:ssä on lisäksi toinen merkittävä puute hyperlinkkien käsitteilyssä: HTML-tiedostoon tallennetaan konversiossa linkin väreiksi Wordissa oletuksena olevat sininen ja punainen. Lisäksi linkkitekstiin tulee mukaan alleviivaus. Kyseiset asetukset riippuvat Wordin hyperlinkki-tyylin asetuksista, mikäli niitä muutetaan, ovat värit ja muotoilut myös HTML-dokumentissa erilaiset. Word 97:n HTML-muunnos yrittää siis parhaansa mukaan saada HTML-dokumentin saman näköiseksi kuin alkuperäinen Word-dokumentti. Tuloksena on kuitenkin alkuperäisen HTML-filosofian¹ vastaista HTML-koodia.

Silloin kun HTML-dokumentista poistetaan turhia linkkejä joudutaan siis linkkitekstistä poistamaan erikseen lisäksi väärä väri ja alleviivaus (tai muut muotoilut).

¹HTML-tiedoston on tarkoitus sisältää vain dokumentin rakenne, ulkomuodosta vastatkoon WWW-selain.

8 Kuvat

Pagemakeria käyttänyt odottaa varmaankin, että myös Word97 osaa muuttaa dokumentin kuvituksen automaattisesti HTML-tiedostoihin sopivaksi. Aivan automaattisesti kuvien muunto ei onnistu, mutta jotain Word97 osaa kyllä tehdä.

HTML-tiedostot eivät sisällä itse kuvia, vaan ne on talletettu erillisiin tiedostoihin, joista HTML-sivulta on linkki. Kuvien tiedostomuodoissa pitää ottaa lisäksi huomioon selainten rajoitukset. GIF- ja JPEG-tiedostoja voidaan käyttää huoletti, lisäksi uusimmissa selaimissa myös PNG-kuvia.

Word-dokumenteissa kuvat voivat olla joko erillisinä tiedostoina tai upotettuna itse dokumenttiedostoon. Lisäksi Wordilla voidaan myös piirtää kaavioita, jotka koostuvat useista irrallisista elementeistä.

8.1 Wordillä piirretyt kaaviot

Muistin loppuraportti sisälsi yhden kaavion, jonka olin tehnyt Wordin piirrustustoiminnoilla. Kaavio itseasiassa koostui irrallisista tekstikehyksistä ja nuolista. Se ei siis muodostanut mitään kuvatiedostoa, jonka olisin voinut koettaa tallettaa johonkin grafiikkamuotoon ja yrittää sitten muuntaa GIF-kuvaksi.

Wordin HTML-konversio jätti kaavion pois kokonaan. Tein kuitenkin hieman kokeiluja ja onnistuin lopulta lisäämään kaavion HTML-tiedostoon GIF-kuvana. Wordin tekemässä HTML-tiedostossa oli ainoastaan jäljellä kaavios- ta sen kuvateksti. Jatkoin HTML-tiedoston käsittelyä Wordilla ja kokeilin lisätä alkuperäisen kuva2.doc-tiedoston, joka sisälsi raportista sen sivun, jossa kaavio oli, objektina HTML-tiedostoon. (Valikosta ”Lisää” valitaan kohta ”objekti” ja esiin tulevasta valintaikkunasta valitaan kohta ”tiedosto” ja etsitään oikea tiedosto.) Mielenkiintoista kyllä lisäys onnistui. Seuraavaksi päätin muuttaa objektin Word-dokumentista, Word-kuvaksi. Klikkasin objektia hiiren oikealla näppäimellä ja esiin tuli valikko, jossa oli kohta ”objekti” ja siitä tuli alavalikko, josta valitsin kohdan ”muunna”. Esiin tuli ikkuna josta sai valita muunnon lopputulokseksi Word-kuvan. Muunnoksen jälkeen paljastui, että minulla oli HTML-tiedosto ja siinä kaavio GIF-kuvana. Word97 oli yllättäen jossain välissä tehnyt konversion automaattisesti.

Loppujen lopuksi jouduin toistamaan operaation muutaman kerran uudelleen, sillä kuva hieman liikaa vasemmalla ja siitä leikkautui osa pois. Vasemman marginaalin liikuttelu ei auttanut, joten jouduin avamaan uudelleen kuva2.doc-tiedoston ja korjailemaan kuvaa Wordissä. Lisäksi HTML-konversion lisäämä kuvateksti piti poistaa, sillä sama kuvateksti oli nyt myös GIF-kuvassa.

Hienoa sinänsä, että kuvan muuntaminen onnistui. Operaatio tuntuu kyl-
lä kokemattomalle käyttäjälle turhan mutkikkaalta. Täytyy kuitenkin pitää
mielessä, että objekteina voi aina koettaa lisätä muita tiedostoja, jos mikään
muu ei auta.

9 Taulukot

Mikäli taulukko on tehty Wordissa sen taulukko-operaatioilla, ei HTML-
versiossa pitäisi olla ongelmia. Aivan kaikkia Wordissa mahdollisia muotoiluja
tosin voida esittää HTML-koodilla.

Word määrittelee taulukon koon pikselin tarkkuudella, jotta konversiossa
käyttöön saataisiin suhteelliset prosenttiarvot, pitää tehdä muutos Wordin
rekisteritietoihin.² Todella käyttäjäystävällistä!

10 Sisällysluettelo

Wordin automaattisesti generoituva sisällysluettelon viimeisin versio siirtyy
HTML-dokumenttiin, luonnollisesti ilman sivujen numeroita. Wordin ohje
väittää, että sivunumeroiden tilalle tulee linkki, joka johtaa vastaavaan koh-
taan tekstissä. Kokeiltavissa dokumenteissa sisällysluettelo oli kuitenkin teh-
ty käsin, jolloin se siirtyi sellaisenaan, tavallisena tekstinä.

Samalla tavalla toimivat automaattisesti generoituva kuvaluettelo, taulu-
koiden luettelo ja hakemisto.

11 alaviitteet

Wordin HTML-konversio poistaa kaikki alaviitteet. Nämä olisi sentään voinut
esittää esimerkiksi omalla sivullaan, johon tekstistä olisi tarvittaessa linkki.
Ilmeisesti liian kova pala Wordin ohjelmoijille tai sitten alaviitteitä ei juuri-
kaan käytetä tavallisissa toimistodokumenteissa Amerikassa eli piirre, jota ei
Microsoftin mielestä tarvitse tukea.

12 Ylä- ja alatunnisteet

HTML ei tue ylä- ja alatunnisteita, koska HTML-dokumentissa ei ole mitään
sivutustakaan. Näitä ei siis konvertoida.

²Katso lisätietoja Wordin ohjeista.