Virheet Csv - Calc UTF8 ohjeet

Suomalaisten korkeakoulujen julkaisutoiminta on kansainvälistä. Julkaisujen ja tekijöiden nimissä esiintyy merkkejä, jotka eivät mahdu yhteen perinteiseen maakohtaiseen merkistöön. Kerättyjen tietojen automaattinen käsittely johtaa virheellisiin tuloksiin, jos osa merkeistä rikkoontuu tiedon-siirrossa. Globalisoituneen maailman merkistö- ja tiedonsiirtostandardit on luotu jo yli 20 vuotta sitten, mutta kaikki ohjelmistot (mm. Excel) eivät edelleenkään tue näitä standardeja.

Tässä ohjeessa kuvataan yksi tapa miten OKM:n julkaisutiedonkeruun edellyttämä UTF-8-merkistökoodattu CSV-siirtotiedosto voidaan tuottaa. Ohje perustuu ilmaisen Open Office -ohjelmistoon kuuluvaan Calc-taulukkolaskentaohjelmaan. Ohjelma on ladattavissa osoitteesta http://www.openoffice.org/fi/. Korkea-kou-lut ylläpitävät julkaisutietoja erilaisissa tietojärjestelmissä, joten tässä ohjeessa on mahdoton kuvata kaikille korkeakouluille yleispätevällä tavalla tietojen toimittaminen alusta loppuun.

Tämän ohjeen lähtökohta on, että OKM:n julkaisutiedonkeruun edellyttämät tiedot luodaan Calc-ohjelmalla tai ladataan siihen (File-valikon Open-toiminnolla) olemassa olevasta tiedostosta. Ensimmäiseksi on syytä varmistaa silmämääräisesti, että tiedot näyttävät kaikin puolin oikeellisilta (esim. erikoismerkit näkyvät oikein ja tiedot ovat oikeissa sarakkeissa). Calc-ohjelma huolehtii lainausmerkkien lisäämisestä automaattisesti puolipisteitä sisältäviin tekstikenttiin, joten lainausmerkkejä ei pidä lisätä käsin kenttiin!

blocked URL

Kuva 1. Tiedoston tarkastelu Calc-ohjelmassa.

Jos tiedosto näyttää olevan kunnossa, niin seuraavaksi valitaan File-valikosta Save As ja annetaan tiedostolle nimi, tiedoston tyypiksi valitaan ”Text CSV (. csv)” ja laitetaan rasti ruutuun Edit filter settings kuten kuvassa 2.

blocked URL

Kuva 2. Save As -valikon oikeat valinnat.

Save-näppäimen painamisen jälkeen näytölle tulee varoitus, ettei kaikki muotoilu säily CSV-muodossa. Tätä ei tarvitse hätkähtää, vaan ilmoitus sivuutetaan tyynesti painamalla Keep Current Format -näppäintä.

blocked URL

Kuva 3. Varoitus CSV-muodon muotoilurajoituksista.

Tämän jälkeen näytölle ilmestyy Export Text File -ikkuna (kuva 4), jossa vetovalikoista valitaan merkistöksi Unicode (UTF-8), kentän erottimeksi puolipiste (;) ja tekstierottimeksi lainausmerkki (”). Näiden valintojen jälkeen painetaan OK-näppäintä.

blocked URL

Kuva 4. Export Text File -valintaikkunan oikeat valinnat.

HUOM! Excel ei osa näyttää syntynyttä CSV-tiedostoa enää oikein (erikoismerkit korruptoituvat), joten jos haluaa varmistua lopputuloksen onnistumisesta syntynyt CSV-tiedosto pitää avata esim. Notepad-ohjel-mal-la, joka osaa näyttää UTF-8-merkistöä.

blocked URL

Kuva 5. Syntyneen CSV-tiedoston tarkastelu Notepad-ohjelmalla.

Tiedosto on nyt valmis ladattavaksi KOTAan.

HUOM.

Puolipiste ja lainausmerkki ovat CSV-formaatin erotinmerkkejä, joita tiedostoa käsittelevä tietokoneohjelma käyttää jakaessaan tiedoston sarakkeiksi ja riveiksi. Julkaisun viitetiedoissa nämä merkit ovat varsin tavallisia ja useimmissa tapauksissa Calc osaakin tulkita soluun kirjoitetut puolipisteet ja lainausmerkit oikein ilman tiedoston rakenteeseen liittyvää erityismerkitystä. Joissain harvoissa tapauksissa - etenkin jos soluun on kirjoitettu pariton määrä lainausmerkkejä saattaa Calc tehdä virhetulkinnan, joka johtaa tiedoston jäsentymiseen väärin ja tallentaa tiedon väärässä muodossa.

Yksi tapa paikallistaa ja korjata näitä virhetulkinnan mahdollisuuksia on korvata ns. ASCII-lainausmerkit kaarevilla kokolainausmerkeillä etsi-korvaa -toiminnolla (Edit-valikosta valitaan Find & Replace). Kaarevat kokolainausmerkit löytyvät Insert-valikon alta Special Character -ikkunasta.

blocked URL

Kuva 6. ASCII-lainausmerkin korvaaminen kaarevilla kokolanausmerkeillä.

Vanha sivu: Suorien tiedonkeruiden (KOTA) ohje UTF-8-merkistökoodatun CSV-tiedoston tuottamiseksi Calc-ohjelmalla.