Konversio-ohjelman lokitiedostot

https://github.com/NatLibFi/yso-marcbib/blob/master/README.md

Konversio-ohjelma tuottaa useita lokitiedostoja, joissa on kussakin hieman eri formaatti. Kenttien erotusmerkkinä on käytetty pystyviivaa "|".

Nimet ovat muotoa

yso-konversio_error-log_VVVV-KK-PPTHHMMSS.csv
yso-konversio_new-fields-log_VVVV-KK-PPTHHMMSS.csv
yso-konversio_removed-fields-log_VVVV-KK-PPTHHMMSS.csv
yso-konversio_results-log_VVVV-KK-PPTHHMMSS.log

error_log

- Tarkistuslista kentistä, joita ei konvertoitu yso- tai slm-termeiksi syystä tai toisesta
- Listaus sisältää 6 saraketta:
  - virhetyyppi
    - 1 - ei löytynyt sanastosta, viety 653:een
    - 2-3-4 - termille useita vaihtoehtoja, termi jätetty paikalleen, sanastotunnus poistettu, toiseen indikaattoriin 4
    - 6 - termi poistettu kokonaan, eri syistä (asiasanana fiktio, aiheet, musiikki, asiasanaketjussa $e-osakenttä tai tyhjä osakenttä)
    - 7 - 650- tai 651-kentässä $g-osakenttä (ei ole varsinainen asiasanakenttä, termi siirretty 653-kenttään)
    - 8 - MARC-formaattiin kuulumaton osakenttätunnus tai kenttä ei sisällä asiasanakenttiä
    - 9 - kenttä sisältää osakentän $6, translitteroidut termit, sanastotunnus poistettu ja 2. indikaattori 4, muuten jätetty paikalleen
  - melinda-tietueen id
  - konversion tyyppi - (kertoo millä konversiosäännön ehdoilla ohjelma on käsitellyt tietueen)
    - m - musiikkiaineisto
    - e - elokuva-aineisto
    - f - fiktio ja pelit
    - t - tietokirjallisuus ja kaikki muu aineisto
  - käsitelty termi
  - alkuperäinen kenttä
  - konvertoitu kenttä
- Esimerkkejä tarkistuslistasta
  - `1|000143880|t|Asia|=650  \7$aluonto$zAsia$2ysa|=653  \5$aAsia`
  - `2|000279076|t|skydd|=650  \7$askydd$2allars|=650  \4$askydd`
  - `3|000144262|t|arvostelu|=650  \7$akirjallisuus$xarvostelu$zAfrikka$2ysa|=650  \4$aarvostelu`
  - `4|1162542|t|mallit|=650  \7$amallit$2ysa|=650  \4$amallit`
  - `6|000143745|t|musiikki|=650  \7$ateatteri$xmusiikki$zSuomi$2ysa`
  - `8|000278142|t|Meksiko|=650  \4$amatkakuvaukset$uMeksiko|=650  \4$amatkakuvaukset$uMeksiko`
  - `9|000306274|t|880-05|=651  \7$6880-05$aCelabinskaa oblast'$2ysa|=651  \4$6880-05$aCelabinskaa oblast'`

Tiedoston käsittely excelissä. Sarakkeita lajittelmalla ja fillteröimällä voi tutkia osajoukkoja. Pivot tablen avulla voi tuottaa ristiintaulukointia ja tilastontia

Avaa suoraan exceliin tai Excelissä poimi data komennolla Data / Get external data / From text ja valitse Delimited

1|000143880|t|Asia|=650 \7$aluonto$zAsia$2ysa\|=653 \5$aAsia						

Jos luet tiedoston sisään sellaisenaan, sen voi pilkkoa sarakkeiksi valitsemalla komentopalkista Data/Text to columns ja valitsemalla Delimited sekä merkillä "|" erotellut sarakkeet.

Kannattaa merkitä kaikki tekstimuotoon, etteivät nollat häviä.

Muista valita merkistöksi UTF-8.

1	000143880	t	Asia	=650 \7$aluonto$zAsia$2ysa	=653 \5$aAsia

Tähän kannattaa pivot tablea varten vielä poimia kentän numero omaksi sarakkeekseen samalla tavalla kuin äsken

kaksi tapaa. Voit kopioida pilkottavan sarakkeen, antaa text-to-columns komennon ja valita tällä kertaa lukutavaksi "Fixed width".

Merkitse raja ja valitse vain säilytettävä sarake, jolloin pilkottu osa jää paikalleen. Valitse muiden sarakkeiden kohdalla "Do not import column (skip)

Vaihtoehtoisesti voit lisätä tyhjän sarakkeen ja poimia tekstin alkuosan viereisestä sarakkeesta funktiolla =LEFT(F1;8)

1	000143880	t	Asia	=LEFT(F1;8)	=650 \7$aluonto$zAsia$2ysa	=LEFT(F1;8)	=653 \5$aAsia

Sen jälkeen rivi näyttää tältä. indikaattorit voi erottaa kenttänumerosta samalla tavalla omaksi sarakkeekseen.

1	000143880	t	Asia	=650 \7	=650 \7$aluonto$zAsia$2ysa	=653 \5	=653 \5$aAsia	

Pivot tablen voi luoda kohdasta Insert / Pivot table

new_fields_log

- Kaikki ohjelman kirjoittamat uudet yso- ja slm-kentät sellaisenaan
- Rivit sislätävät kolme "|" merkillä erotettua kenttää: Melinda-id, konversiotyyppi, kirjoitettu kenttä
- Esimerkki
    - ```
      1386723|m|=370 \\$81\u$gSaksa$2yso/fin$0http://www.yso.fi/onto/yso/p105087
      1386723|m|=382 11$81\u$asello$2seko
      1386723|m|=388 \\$81\u$a1720-luku$2yso/fin
      1386723|m|=655 \7$81\u$asarjat$2slm/fin$0http://urn.fi/URN:NBN:fi:au:slm:s887
      ```
    - ```
      Musiikkiaineiston ketjun osakentät on purettu omiin kenttiinsä. Samaan ketjuun kuuluneiden termien kentät on merkitty $8 osakentällä ja ketjun järjestysnumerolla.
      ```

removed_fields_log

- Kaikki ohjelman poistamat kentät
- Rivit sisältävat kaksi saraketta "|" merkillä erotettua kenttää: Melinda-ID, alkuperäinen kenttä
- Esimerkki: (kts vastaava new fields esimerkki)
    - ```
      1386723|=650  \7$asarjat$xsello$zSaksa$y1720-luku$2musa
      ```

results_log

- Raportti ohjelman käsittelemistä rivimääristä
- Esimerkki raportista:
- ```
  konvertoituja tietueita: 999931
  käsiteltyjä tietueita: 999979
  käsiteltyjä kenttiä: 13310430
  kaikki tarkistetut kentät: 4860901
  poistettuja kenttiä: 4208286
  uusia kenttiä: 9102144
  MARC21-virheitä: 1
  Virhetilastot:
  Virhetyyppi: UnicodeDecodeError, määrä: 1
  ```
- Mikäli raportissa löytyy MARC21 virheitä, niin lähtöaineistossa voi olla virheitä, jotka olisi hyvä korjata ensin, jotta konversio onnistuu.