OCR-aineistojen editointi Sukukielten digitointiprojektissa

Jussi-Pekka Hakkarainen
Projektipäällikkö
Kansalliskirjasto - Tutkimuskirjasto

5.6.2013, Variantti-kollokvio, SKS, Juhlasali

Sukukielten digitointiprojekti

- **Sukukielten digitointiprojekti – pilotointi 07/2012–10/2013**
  - Koneen Säätiön Kieliohjelman osahanke.
  - Tuotetaan Kieliohjelman tutkimukselle tutkimusaineistoa ja –infrastruktuuri.

- **Digitoinnit julkaistu Fenno-Ugrica –kokoelmassa**
  - 17 000 sivua / 131 monografiaa, inkeroisten-, vepsän, marin- (niittymari ja vuorimari) ja mordvalaiskielillä (ersä ja mokša). Monografiat pääsääntöisesti koulu- ja oppikirjoja 1920-1930-luvuilta.
  - 20 000 sivua marilaisia ja mordvalaisia sanomalehtiä.
  - Aineisto digitoitu Venäjän Kansalliskirjaston (Pietari) kokoelmista.
  - Tekijänoikeuksien selvitys yhdessä venäläisten toimijoiden kanssa.

OCR-editori

- **Sukukielten digitointiprojektissa kehitetty OCR-editori**
  - Tarkoitus korjata digitoinnissa jääneitä virheitä
  - Voidaan lisätä puuttuvia kirjainmerkkejä (UTF-8)
  - Ei integroitu julkaisuarkistoon, vaan oma käyttöliittymä
  - Vapaan lähdekoodin ohjelmisto

- **Etuja ja puutteita**
  - Voidaan korjata toistuvia virheitä
  - Käyttäjähierarkia
  - Editoidut aineistot - .xml-, .txt- ja .pdf-formaateissa
  - PDF:n/OCR:n koordinaatit erittäin hankalasti hallittavissa
  - Palstoituksen tekeminen aineistoihin pulmallista

Yhteystiedot

Projektin kotisivu
http://www.kansalliskirjasto.fi/kokoelmatjapalvelut/digitaalisetkokoelmat/finnougric.html

Fenno-Ugrica -kokoelma
fennougrica.kansalliskirjasto.fi/

Projektiblogi
https://blogs.helsinki.fi/fennougrica/

Palveluosoite
kk-fennougrica@helsinki.fi

Jussi-Pekka Hakkarainen
Kansalliskirjasto
PL 26 (Teollisuuskatu 23)
00014 Helsingin yliopisto
jussi-pekka.hakkarainen@helsinki.fi