

Tuomiokirjahaku ja Transkribus ym.

Kuinka sukututkija voi hyödyntää erilaisia käsinkirjoitetun tekstin tunnistustyökaluja

Kari Kujansuu
Tampereen Seudun Sukututkimusseura
23.9.2025

Tarve

- yksittäinen dokumentti tai muutama kuva selväkieliseksi (pöytäkirja, kirje, akti, ...)
 - KA Tekstintunnistustyökalu
 - Transkribus
 - tekoälypalvelut (ChatGPT, Google Gemini etc)
 - vaikka tekstin pystyisi helposti itsekin lukemaan, niin nopeuttaa puhtaaksikirjoitusta
- suurempi massa: kokonaisen kirjan (tai arkistosarjan tms) tulkitseminen jotta siihen voisi tehdä hakuja
 - tuomiokirjat.kansallisarkisto.fi
 - Transkribus

Termejä

- OCR – Optical Character Recognition
 - painetulle tekstile
- HTR – Handwritten Text Recognition
 - käsinkirjoitetulle tekstile
- malli (käsialamalli, model)
 - tiettyyn kieleen, kirjoitustapaan tai aikakauteen tms. erikoistunut ”resepti”, joka tunnistaa kirjaimet, sanat, rivit, tekstialueet ym.
- transkriptio, transkribointi, litterointi – kuvana olevan tekstin tulkinta selväkieliseksi
- CER – Character Error Rate – virheprosentti, mielellään alle 5%

Esimerkkejä tunnistettavasta tekstistä

- Tuomiokirja, Hollola 1653
- Ylöjärvi muuttaneet 1862
- Henkikirja 1870
- Henkikirja 1910
- kirje 1960

Palveluja

- Kansallisarkiston tekstintunnistustyökalu
- Transkribus
- valmiiksi tulkittuja asiakirjoja
 - KA tuomiokirjahaku: tuomiokirjat 1800-luvulta
 - Ruotsi: Riksarkivet, ArkivDigital
 - FamilySearch
- Microsoft, Google, Amazon...
- tekoälypalvelut (ChatGPT, Copilot, Google Gemini, ...)

KA:n Tekstintunnistustyökalu

- ”Huggingface”
- <https://kansallisarkisto.fi> > Tutki aineistoja > Kaikki verkkopalvelut ja tietokannat > Tekstintunnistustyökalu
- yksi kuva kerrallaan
- maksimikoko?
- toimii 1600-1900-lukujen teksteille
- vastaava myös ainakin Ruotsin Riksarkivetilla

KA:n tuomiokirjahaku

- <https://tuomiokirjat.kansallisarkisto.fi>
- 1800-luvun tuomiokirjoja (varsinaiset ja ilmoitusasiat)
- aikaraja n. 1810-1870
- haku koko materiaalista tai rajaten
- kihlakunnanoikeudet ja raastuvanoikeudet
 - ei luovutettua aluetta, esim Viipuri
 - puuttuu myös Lappeenranta, Hamina
 - ei hovioikeuksia
 - ei linkkiä Astiaan

Transkribus

- transkribus.eu, app.transkribus.org
- READ-hanke/EU, READ-COOP-osuuskunta
- maksullinen, n. 0,25 euroa/sivu, 50 sivua/kk ilmaiseksi
- runsaasti malleja eri kielille ja ajanjaksoille ym
 - Suomi 1870-1917
 - NAF Court Records M10 v2
 - myös painetulle tekstille (halvempi)
 - voi kouluttaa omia malleja

Muuta

- Turun yliopiston tutkimushanke muuttaneiden luetteloiden tulkitsemiseen
 - <https://openhumanitiesdata.metajnl.com/articles/10.5334/johd.345>
 - tulokset avoimena datana: <https://zenodo.org/records/15606656>

Linkkejä

- <https://kansallisarkisto.fi/-/kokeile-uutta-ilmaista-tekstintunnistustyokalua>
- <https://tuomiokirjat.kansallisarkisto.fi/#/>
- <https://chatgpt.com>
- <https://gemini.google.com/>
- <https://www.transkribus.org>
- <https://app.transkribus.org>
- <https://riksarkivet.se/utveckla-och-samarbeta/ai-och-arkiv>
-