Mihin annotointia tarvitaan tekoälyaikana

Tekoälyllä lisäarvoa digiarkistojen asiakkaille
–hankkeen loppuseminaari 28.8.2024

Euroopan unionin osarahoittama

XAMK
Kaakkois-Suomen ammattikorkeakoulu

KANSALLISARKISTO

elka SUOMEN ELINKEINOELÄMÄN KESKUSARKISTO

Taustaa

- Elka on osallistunut Tekoälyllä lisäarvoa digiarkistojen asiakkaille –hankkeeseen osatoteuttajana
- Hankkeen tavoitteet:
    - Tietojen automaattisen poiminnan edistäminen
    - Asiakirjojen sisällöllisten merkitysten tunnistamisen kehittäminen
    - Sisällönkuvailun lisäarvon konkretisoiminen
- Hankkeen ajaksi Elkaan palkattiin opetusaineiston tuottaja, jonka työ on sisältänyt:
    - Asiakirjojen digitointia eli skannaamista
    - Sähköisten asiakirjojen annotointia

Miksi annotointia?

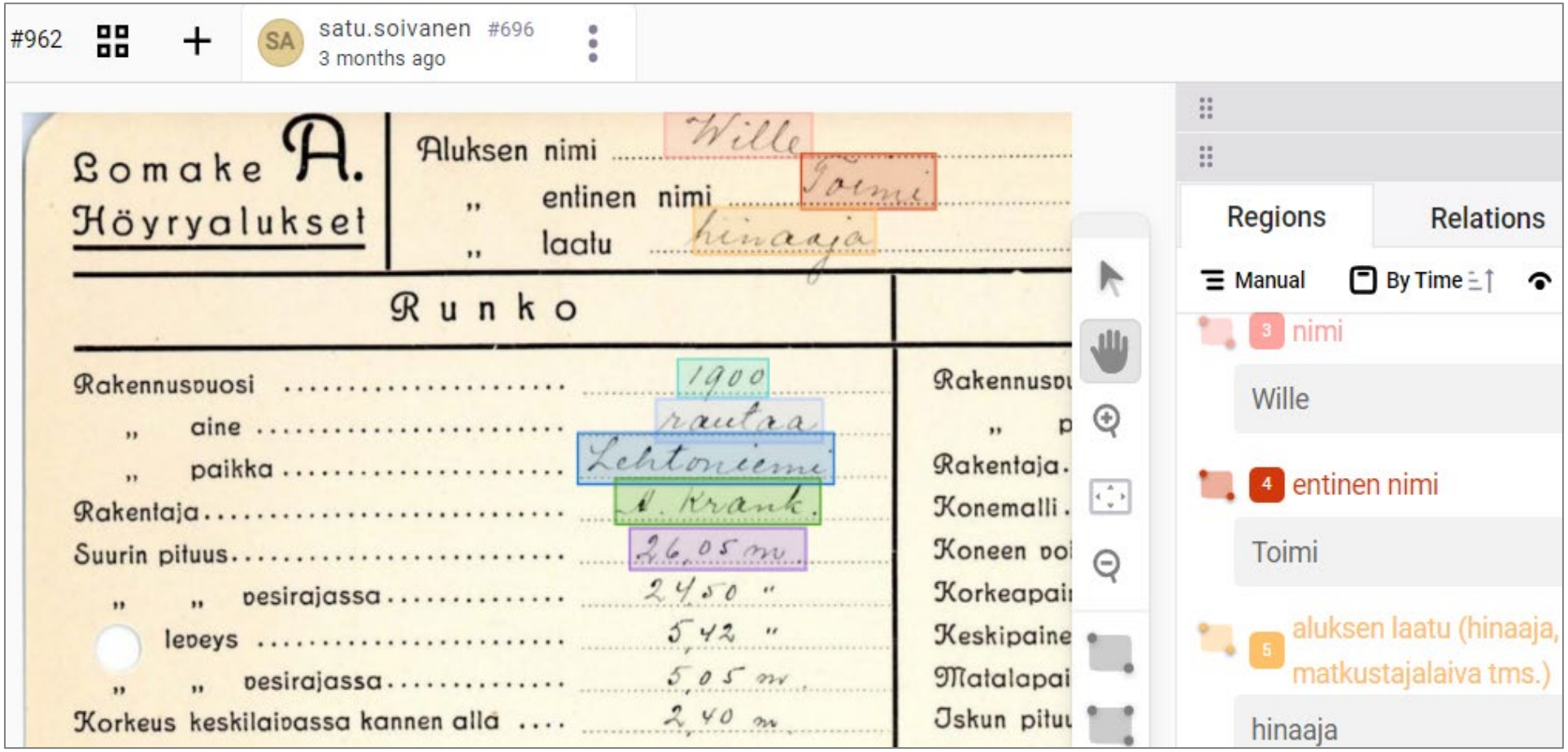
- ChatGPT-4 vastaa kysymykseen ”Mihin annotointia tarvitaan tekoälyaikana?” mm. näin:
  - Annotointi on kriittinen osa tekoälykehitystä ja sen jatkuvaa parantamista.
  - Se varmistaa, että mallit ovat tarkkoja, luotettavia ja toimivat tehokkaasti erilaisissa sovelluksissa.

- Tekoälyllä lisäarvoa digiarkistojen asiakkaille –hankkeessa
  - Annotointi on ollut tarpeellista suomenkielisten tekstien tekstintunnistuksen edelleen kehittämiseen. Tekstintunnistus luo pohjan asiakirjan tekstin semanttiselle analyysille: mikä on asiakirjan sisältö ja asiayhteys.
  - Annotointia on hyödynnetty myös asiakirjan segmentoinnissa. Segmentointi mallintaa asiakirjaa niin, että menetelmä osaa poimia tekstiä oikeasta kohtaa dokumenttia.

Millaisia Elkan dokumentteja on annotoitu

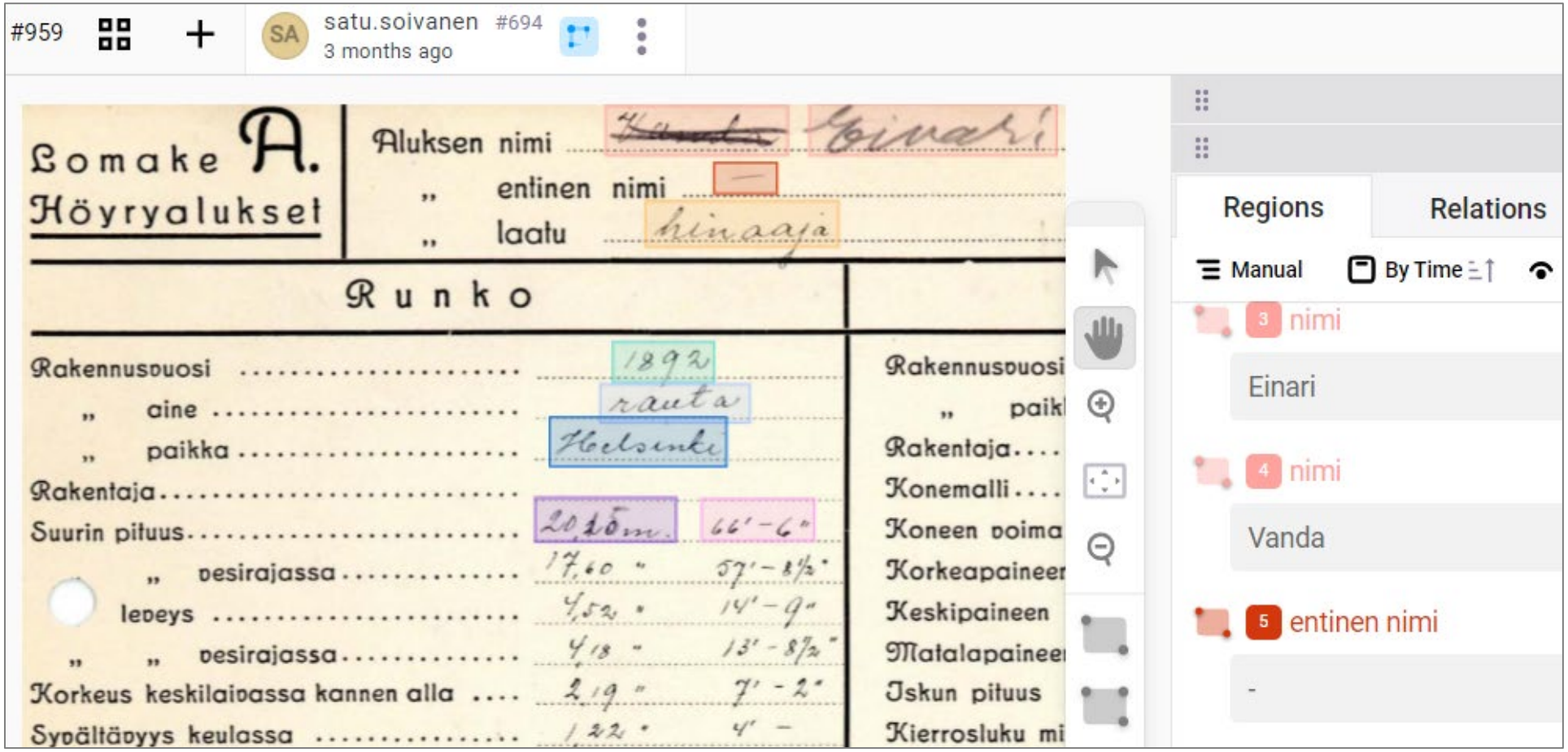
- Hankkeen aikana on Elkan aineistoista digitoitu konekirjoitettuja ja käsinkirjoitettuja asiakirjoja yhteensä noin 36400 sivua, joista on valittu 10 % annotointiin ja testaukseen seuraavilta ajanjaksoilta:
    - Enso-Gutzeit laiva-asiakirjoja 1920-1970
    - Rautaruukki toimintakertomus 1975
    - Riihimäenlasi tuotejulkaisuja 1978-1985
    - Stora Enso toimintakertomuksia 1998-2000
    - YLE vuosikertomuksia 1993-1996
    - YLE kirjeenvaihtoa 1927-1966
- Tekoälyn koulutusaineistona olevasta 3600 sivusta laiva-asiakirjoja on 900 eli neljännes.
- Seuraavana 6 esimerkkiä alusten nimien annotoinnista Label Studio –sovelluksella sekä lopuksi taustaa hankkeessa koulutetusta tekoälymallista ja pari esimerkkiä tuloksista.

Aluksen nimen annotointi 1/6



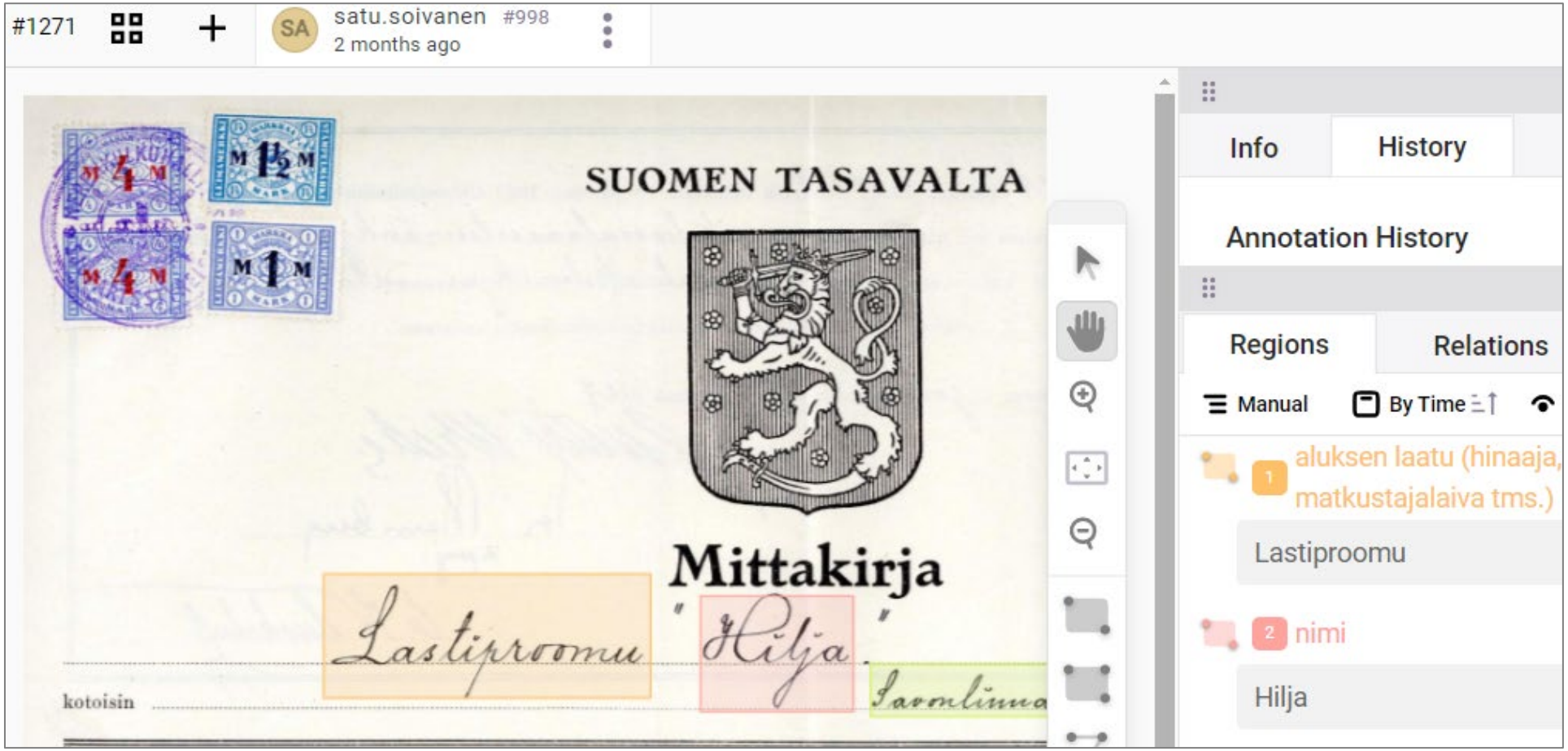
- Aluksen nimi Wille, entinen nimi Toimi

Aluksen nimen annotointi 2/6



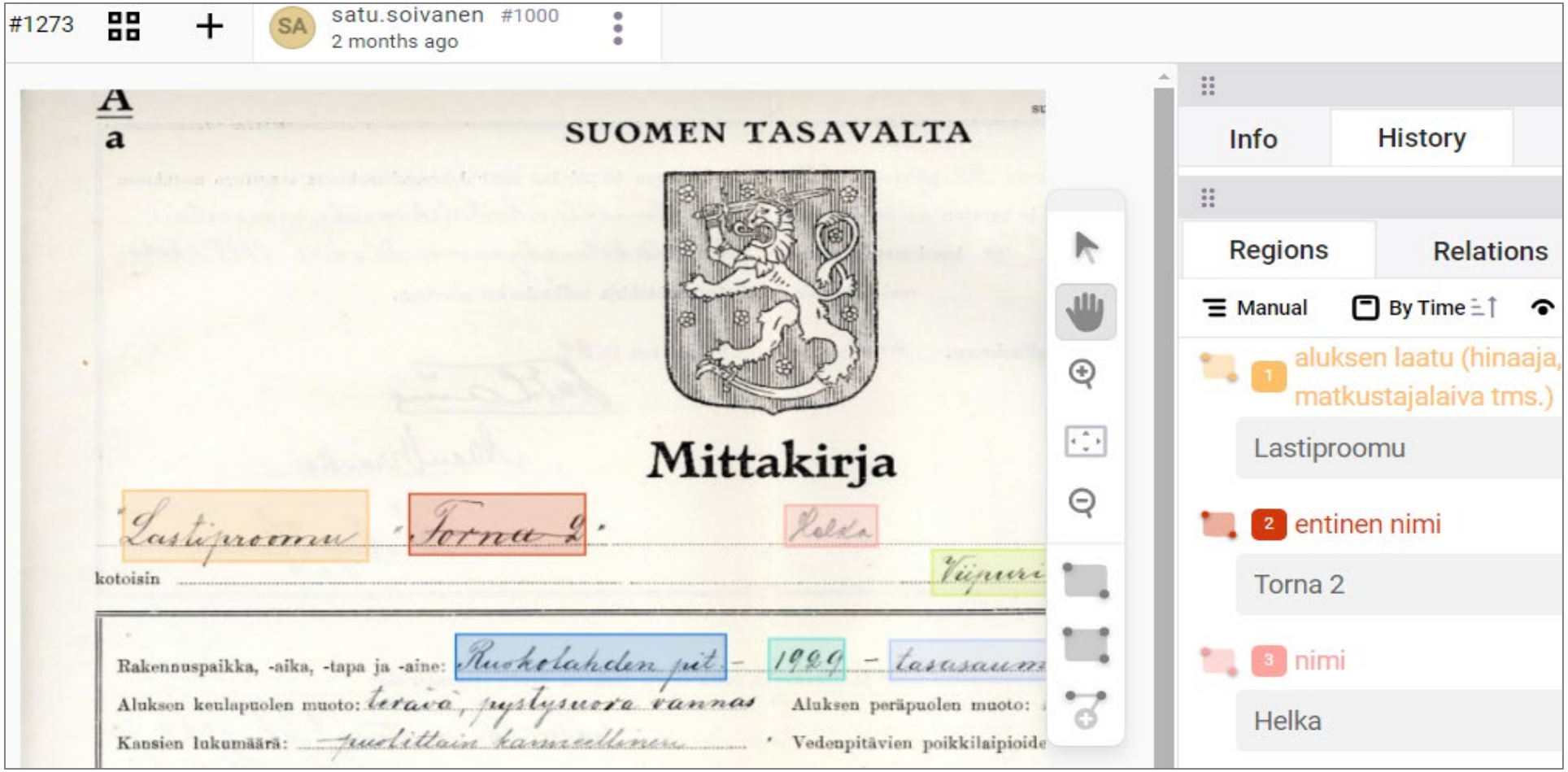
- Aluksen nimi Vanda yliviivattu, uusi nimi Einari (kaksi ilmentymää nimi-luokkaan)
- Aluksen entinen nimi - (viiva)

Aluksen nimen annotointi 3/6



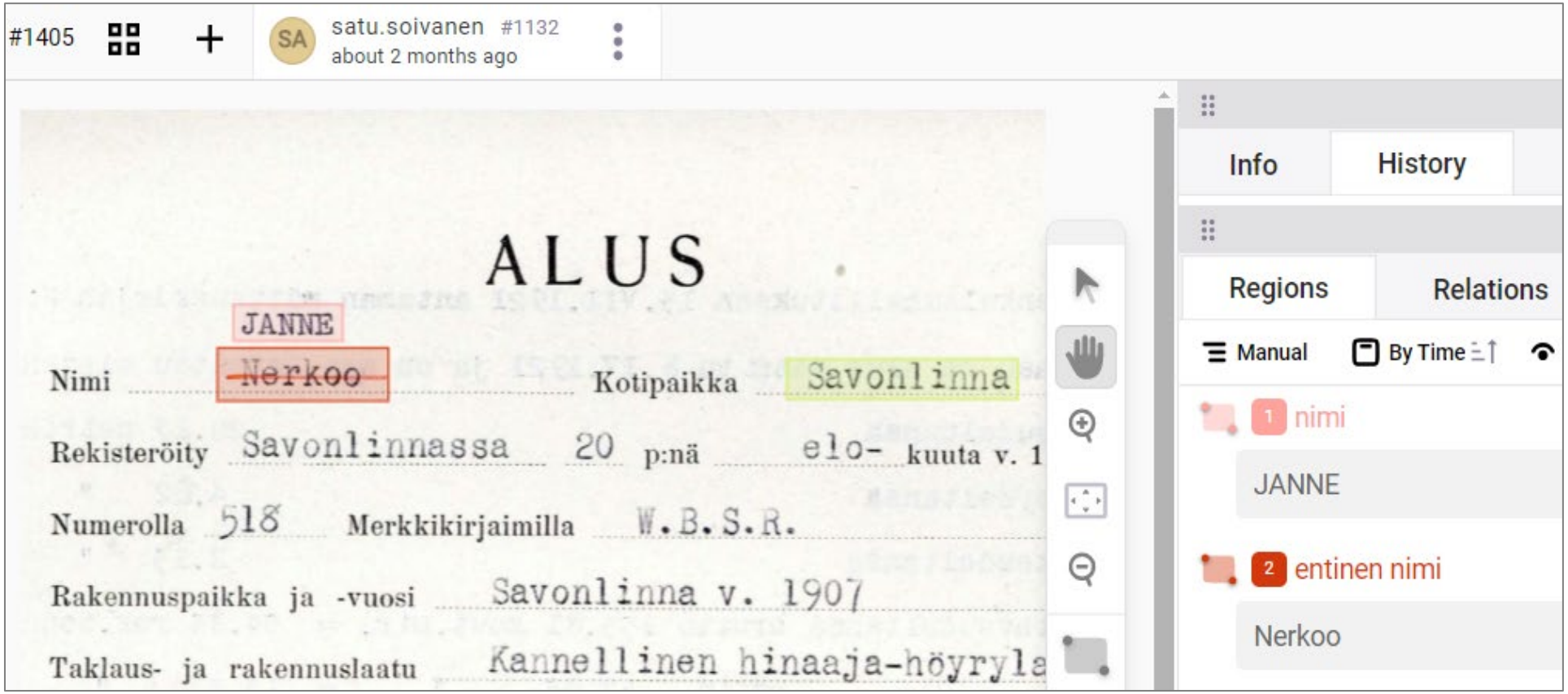
- Aluksen nimi Hilja (lainausmerkkejä ei huomioida), nimen edessä aluksen laatu Lastiproomu

Aluksen nimen annotointi 4/6



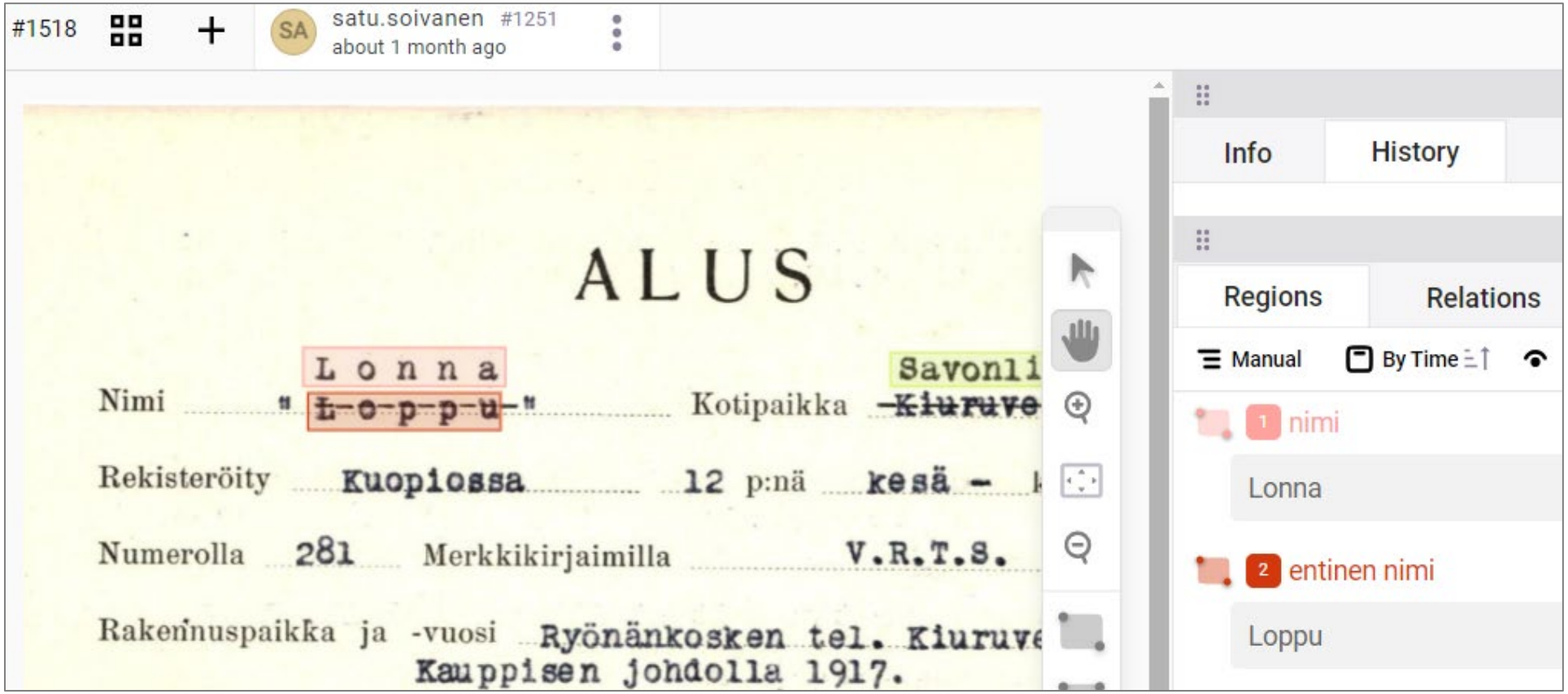
- Aluksen nimi Helka, entinen nimi Torna 2 (yliviivattu), edessä laatu Lastiproomu

Aluksen nimen annotointi 5/6



- Aluksen nimi JANNE (huomioidaan isot kirjaimet), entinen nimi Nerkoo (yliviivattu)

Aluksen nimen annotointi 6/6

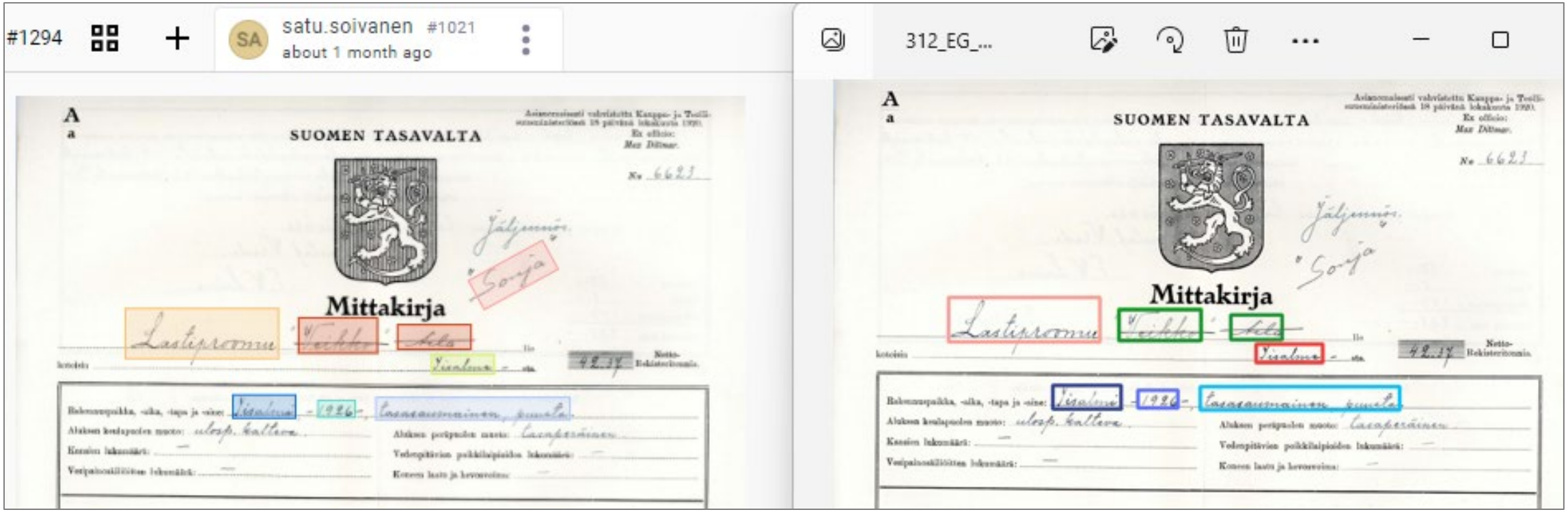


- Aluksen nimi Lonna, entinen nimi Loppu (yliviivattu, ei huomioida välilyöntejä eikä lainausmerkkejä)

Mitä tekoälymallia on opetettu

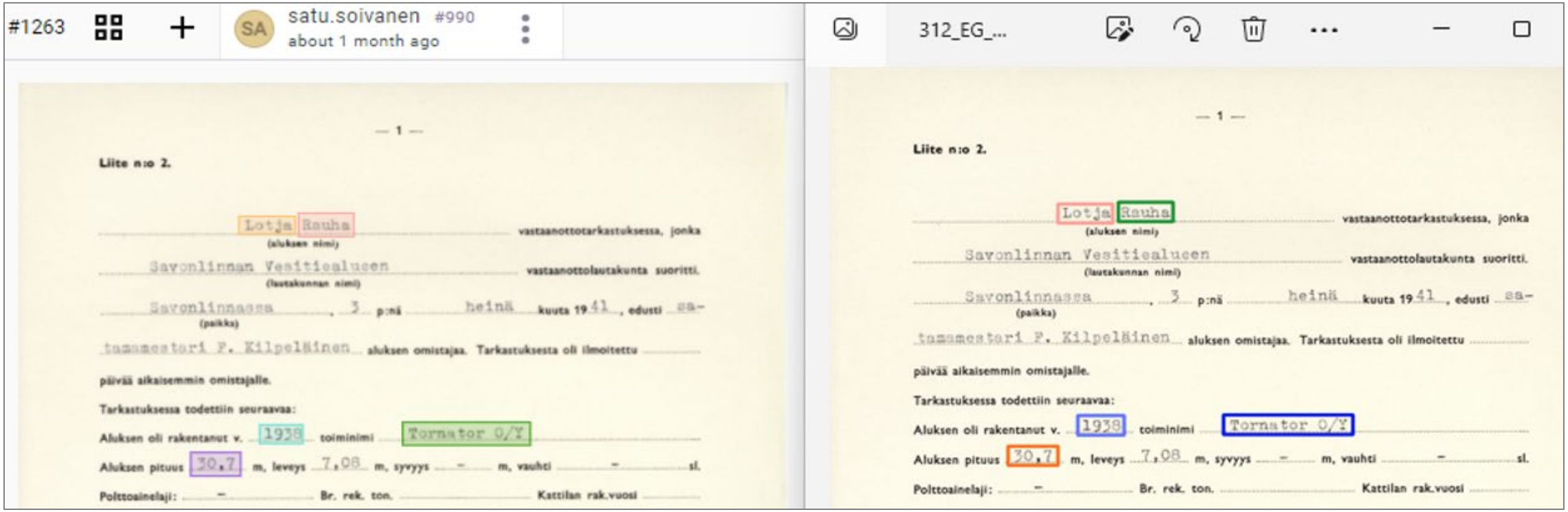
- Hankkeessa on ollut tavoitteena parantaa tekstintunnistuksen laatua käyttämällä tekstintunnistukseen Paddle-OCR:ää
- Mikä on Paddle-OCR? (https://www.plugger.ai/blog/what-is-paddle-ocr)
  - Paddle-OCR on avoimen lähdekoodin menetelmä optiseen merkintunnistukseen.
  - Sen on suunnitellut kiinalainen tekoälyyn ja syväoppimiseen keskittynyt hakukoneyritys, jonka luoma koneoppisen kehitysalusta PaddlePaddle on jo levittäytynyt laajalle.
  - Paddle nimi on lyhenne sanoista PArallel Distributed Deep Learning
  - Paddle-OCR sisältää joukon valmiiksi koulutettuja syväoppimismalleja. Se tunnistaa ja poimii tekstiä erityyppisistä kuvista ja asiakirjoista (mm. JPEG-, PNG-, BMP- ja PDF-muotoisista)
  - Tunnistaa tekstiä jo yli 70 kielellä
- Hankkeessa Paddle-OCR:ää on lisäkoulutettu Elkan ja Kansallisarkiston annotoiduilla aineistoilla.
- Seuraavana lopuksi pari esimerkkiä siitä, mitä Paddle-OCR:llä on jo saatu aikaan.

Aluksen nimen poiminta 1/2



- Kuva vasemmalla: Annotoitu aluksen nimeksi Sonja, entisiksi nimiksi Veikko ja Aila.
- Kuva oikealla: Opetettu Paddle-OCR löytänyt nimet Veikko ja Aila, muttei viistossa olevaa Sonjaa.
- Paddle-OCR osaa poimia käsinkirjoitettua tekstiä melko hyvin oikeasta kohtaa dokumenttia.

Aluksen nimen poiminta 2/2



- Kuva vasemmalla: Annotoitu nimeksi Rauha, kuva oikealla: Paddle-OCR löytänyt nimen Rauha.
- Paddle-OCR osaa poimia konekirjoitettua tekstiä oikeasta kohtaa dokumenttia ko. esimerkissä.
- Konekirjoitettujen asiakirjojen osalta tulokset ovatkin varsin lupaavia.

Kiitos

Satu Soivanen
Opetusaineiston tuottaja

+358 (0)44 321 3412

Suomen Elinkeinoelämän
Keskusarkisto
Tutkijantie 7
50100 Mikkeli

www.elka.fi