

Mihin annointia tarvitaan tekoälyaikana

Tekoälyllä lisäarvoa digiarkistojen asiakkaille
–hankkeen loppuseminaari 28.8.2024



**Euroopan unionin
osarahoittama**



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

#962 SA satu.soivanen #696 3 months ago

Lomake A.
Höyryalukset

Aluksen nimi Wille
" entinen nimi Toimi
" laatu hinaaja

Runko

Rakennusvuosi	1900	Rakennusvuosi	
" aine	rautaa	"	
" paikka	Lehtoniemi	Rakentaja	
Rakentaja	H. Krank.	Konemalli	
Suurin pituus	26,05 m.	Koneen vo	
" " vesirajassa	2450 "	Korkeapain	
" leveys	542 "	Keskipaine	
" " vesirajassa	505 m	Matalapai	
Korkeus keskilaivassa kannen allä	2,40 m	Jskun pitu	

Regions Relations

Manual By Time

3 nimi

Wille

4 entinen nimi

Toimi

5 aluksen laatu (hinaaja, matkustajalaiva tms.)

hinaaja

- Aluksen nimi Wille, entinen nimi Toimi

Aluksen nimen annotointi 2/6

#959 SA satu.soivanen #694 3 months ago

Lomake A.
Höyryalukset

Aluksen nimi ~~Vanda~~ **Einari**
" entinen nimi -
" laatu **hinaaja**

Runko

Rakennusvuosi	1892	Rakennusvuosi
" aine	rauta	" paikka
" paikka	Helsinki	Rakentaja...
Rakentaja		Konemalli...
Suurin pituus	20,10 m. 66'-6"	Koneen voima
" vesirajassa	17,60 " 57'-8 1/2"	Korkeapaineen
leveys	4,52 " 14'-9"	Keskipaineen
" " vesirajassa	4,18 " 13'-8 1/2"	Matalapaineen
Korkeus keskilavassa kannen alla	2,19 " 7'-2"	Jskun pituus
Syvältävyys keulassa	1,22 " 4' -	Kierrosluku mi

Regions Relations
Manual By Time

3 nimi
Einari

4 nimi
Vanda

5 entinen nimi
-

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

Aluksen nimen annotointi 3/6

The image shows a screenshot of a document viewer interface. The document is a Finnish shipping receipt (Mittakirja) from the Republic of Finland (SUOMEN TASAVALTA). It features a coat of arms and the text "Mittakirja". The receipt is annotated with several colored boxes: an orange box around the word "Lastiproomu", a red box around the name "Hilja", and a green box around the word "Saonlinna". The interface includes a top bar with the user's name "satu.soivanen #998" and a date "2 months ago". On the right side, there is a sidebar with tabs for "Info", "History", "Annotation History", "Regions", and "Relations". The "Annotation History" section shows two entries: "1 aluksen laatu (hinaaja, matkustajalaiva tms.)" and "2 nimi".

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

Aluksen nimen annotointi 4/6

#1273 SA satu.soivanen #1000 2 months ago

A
a

SUOMEN TASAVALTA

Mittakirja

Lastiproomu Torna 2 Helka

kotoisin Viipuri

Rakennuspaikka, -aika, -tapa ja -aine: Ruskolahden pit. 1989 Lastusaukko

Aluksen keulapuolen muoto: terävä, pystysuora rannas Aluksen peräpuolen muoto:

Kansien lukumäärä: puuhittain karmasellinen Vedenpitävien poikkilaipioide

Info History

Regions Relations

Manual By Time

1 aluksen laatu (hinaaja, matkustajalaiva tms.)
Lastiproomu

2 entinen nimi
Torna 2

3 nimi
Helka

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

Aluksen nimen annotointi 5/6

#1405 satu.soivanen #1132
about 2 months ago

ALUS

Nimi **JANNE** ~~Nerkoo~~ Kotipaikka **Savonlinna**

Rekisteröity Savonlinnassa 20 p:nä elo- kuuta v. 1

Numerolla 518 Merkkikirjaimilla W.B.S.R.

Rakennuspaikka ja -vuosi Savonlinna v. 1907

Taklaus- ja rakennuslaatu Kannellinen hinaaja-höyryla

Info History

Regions Relations

Manual By Time

1 nimi

JANNE

2 entinen nimi

Nerkoo

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

Aluksen nimen annotointi 6/6

#1518 SA satu.soivanen #1251 about 1 month ago

ALUS

Nimi **Lonna** **Loppu** Kotipaikka **Savonlinna** **Kiuruve**

Rekisteröity **Kuopiossa** 12 p:nä **kesä -**

Numerolla **281** Merkkikirjaimilla **V.R.T.S.**

Rakenuspaikka ja -vuosi **Ryönänkosken tel. Kiuruve**
Kauppisen johdolla 1917.

Info History

Regions Relations

Manual By Time

1 nimi

Lonna

2 entinen nimi

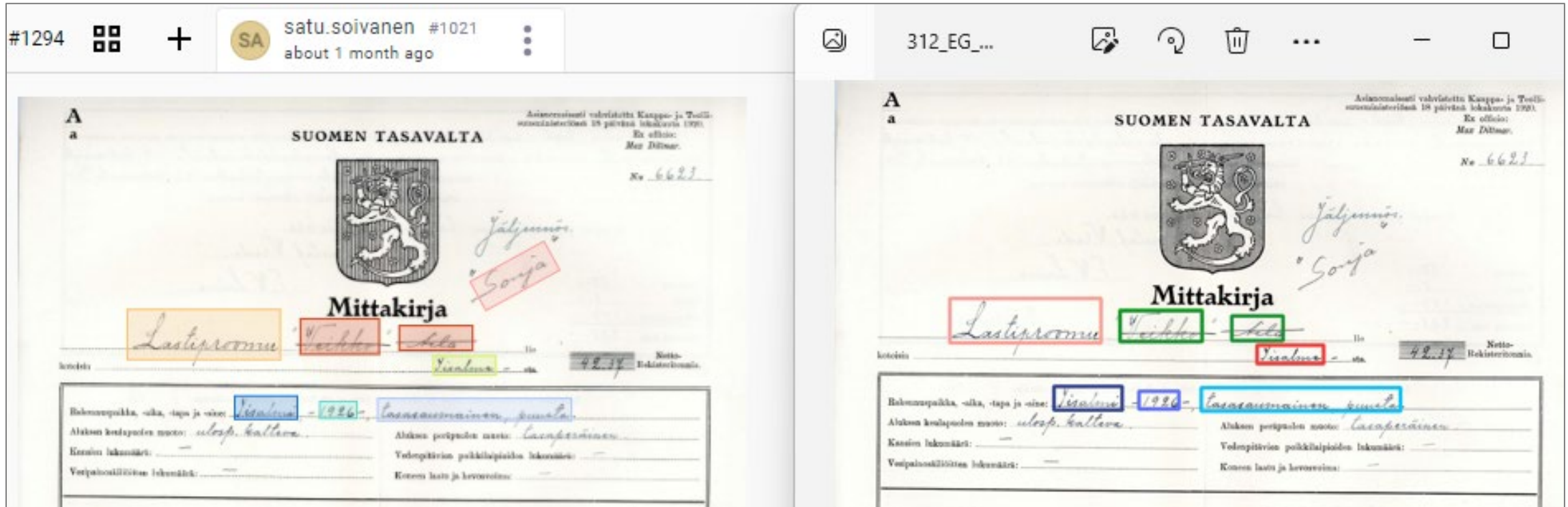
Loppu

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

Mitä tekoälymallia on opetettu

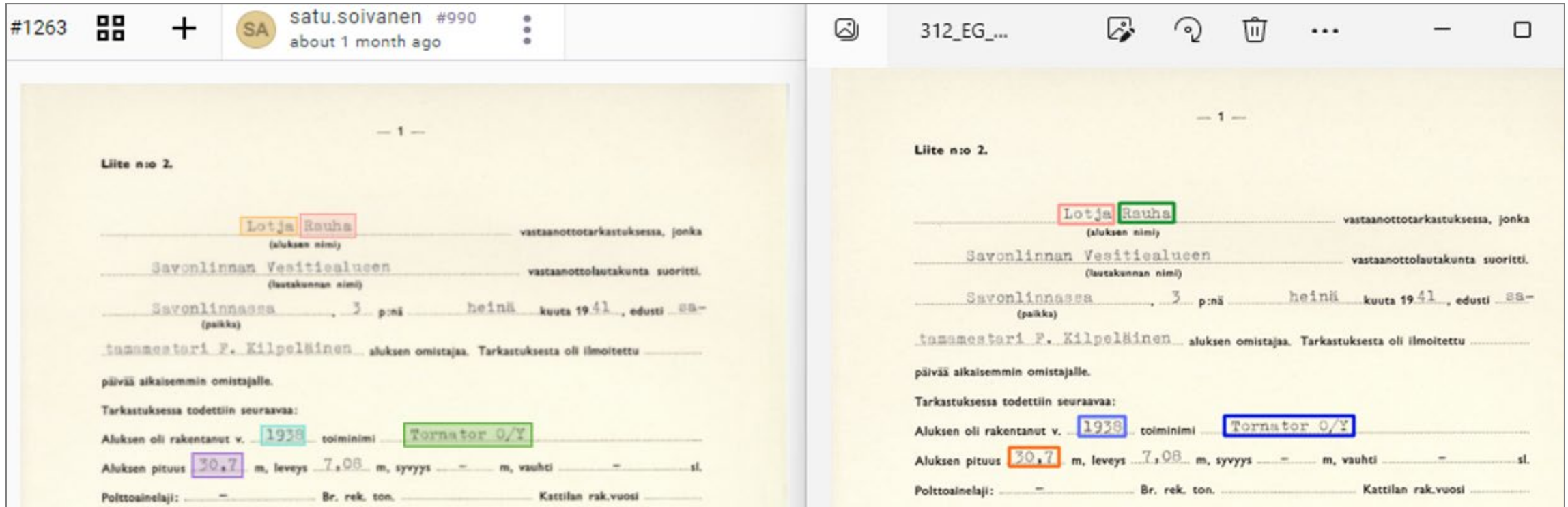
- Hankkeessa on ollut tavoitteena parantaa tekstintunnistuksen laatua käyttämällä tekstintunnistukseen Paddle-OCR:ää
- Mikä on Paddle-OCR? (<https://www.plugger.ai/blog/what-is-paddle-ocr>)
 - Paddle-OCR on avoimen lähdekoodin menetelmä optiseen merkintunnistukseen.
 - Sen on suunnitellut kiinalainen 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



elka

Suomen Elinkeinoelämän
Keskusarkisto
Tutkijantie 7
50100 Mikkeli

www.elka.fi

