



**Euroopan unionin
osarahoittama**

Tekoälyllä lisäarvoa digiarkistojen asiakkaille

Hankkeen päätösseminaari

28.8.2024, 12.30-15.00

Merkki & Teams

Rahoittaja: Etelä-Savon maakuntaliitto, Euroopan
aluekehitysrahasto

Tallenna

Sisältö ja aikataulu

- 12.30 - 12.35 Tilaisuuden avaus, Xamk / Anssi Jääskeläinen
- 12.35 - 12:45 Avauspuheenvuoro, Musiikkiarkisto / Juha Henriksson
- 12.45 - 12.55 [Tekoälyllä lisäarvoa digiarkistojen asiakkaille hanke](#), Xamk / Anssi Jääskeläinen
- 12.55 - 13.55 OCR:n tehostaminen kokonaisuutena
 - 12.55 - 13.05 Mihin annotointia tarvitaan tekoälyaikana, Elka / Satu Soivanen
 - 13.05 - 13.10 Miten annotointi toteutettiin, Kansallisarkisto / Sini Rajaniemi
 - 13.10 - 13.25 Paddle OCR moottorin jatkokouluttaminen annotoiduilla ja synteettisellä aineistolla, Kansallisarkisto / Atte Föhr & Xamk / Tuomo Räisänen
 - 13.25 - 13.35 Paddle OCR pilotit Xamk / Anssi Jääskeläinen
 - 13.35 - 13.45 Testaajan kokemuksia OCR pilotin käytöstä, Merkki / Johanna Mieto
- 13.45 – 14.10 Kahvitauko & Jatkokehitysideoita yhdessä keskustellen, Hanketiimi
 - Etäosallistujat voivat heitellä ideoita chatiin
- 14.10 - 14.30 Fokusryhmähaastattelujen raportin yhteenveto ja refleктоiva keskustelu, Xamk / Mira Kolari & Musiikkiarkisto / Juha Henriksson
- 14.30 - 14.40 Memoriaalin, Arkkiivin ja tämän hankkeen tuloksien yhdistäminen ja hankejatkumo, Xamk / Mira Kolari
- 14.40 - 14.55 Koneoppimista ja RAG hyödyntämistä, Kansallisarkisto / Mikko Lipsanen
- 14.55 - 15.00 Yhteenveto ja päätös, Xamk / Noora Talsi

Käytännön asiat

- Koko tilaisuuden ajan käytössä
 - <https://www.menti.com/aldowixac9jy>
 - Palautetta, kommentteja, kehitysehdotuksia, jne.
- Äänentoisto, pieni Jabra. Kysymykset etäosallistujilta mieluusti chattiin
- Wifi: Museon vieraille/Paivalehti





**Euroopan unionin
osarahoittama**

Tekoälyllä lisäarvoa digiarkistojen asiakkaille

Hanke-esittely

Anssi Jääskeläinen

Anssi.Jaaskelainen@xamk.fi

Rahoittaja: Etelä-Savon maakuntaliitto, Euroopan
aluekehitysrahasto

Hanke ja tavoitteet

- Edistää tekoälyn hyödyntämistä aineistojen löydettävyyden ja jatkokäytön kannalta
- Fokus loppukäyttäjien näkökulmassa
- Edistää tietojen automaattista poimintaa
- Kasvattaa osaamista ja ymmärrystä tekoälyn hyödyntämispotentiaalista
- Tehostaen aineistojen kuvailua
- Kehittää aineistojen sisällöllisten merkitysten tunnistamista

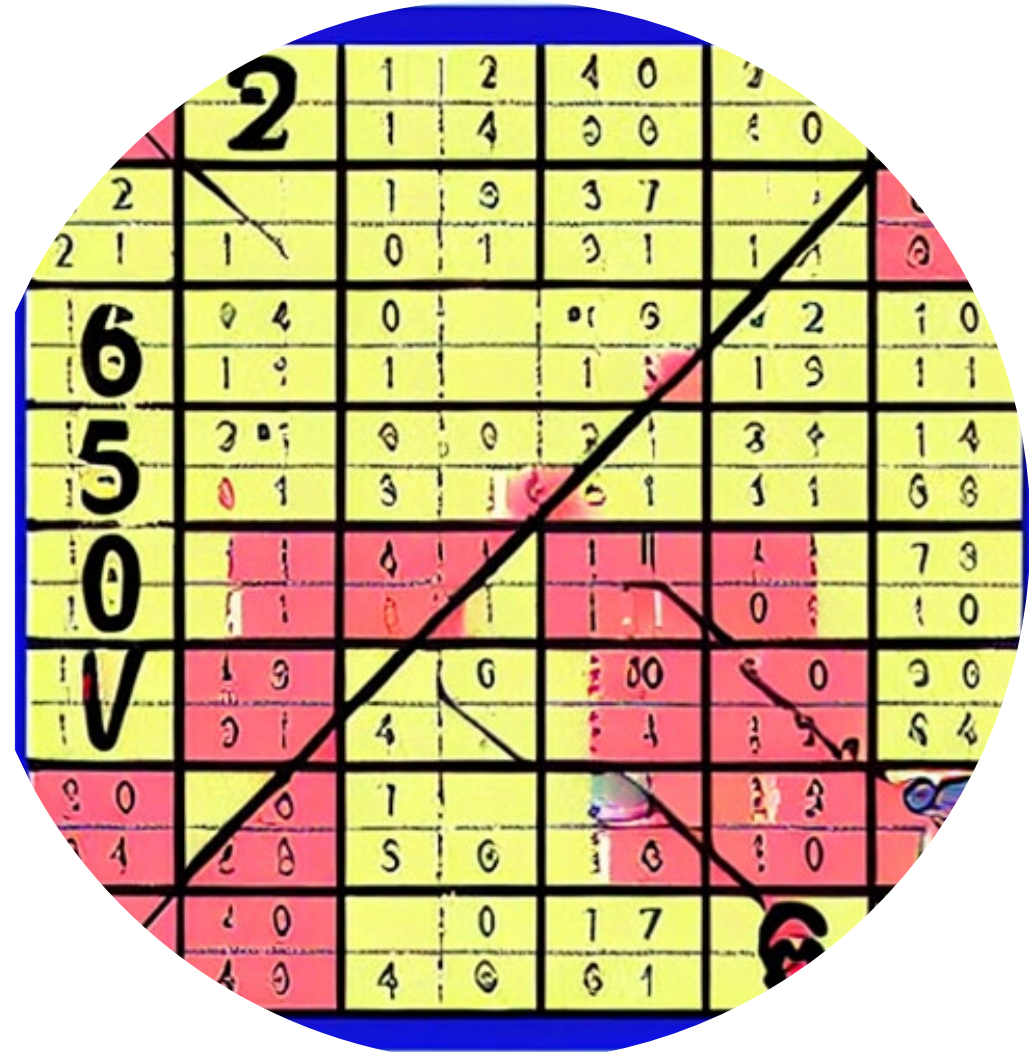


KANSALLISARKISTO



Toimenpiteet

- Tekstintunnistuksen ja sivusegmentoinnin tehostaminen
 - Helpompaa poimia tärkeitä asioita dokumenteista
 - Parempi tunnistustarkkuus
 - Kokeiltiin myös kielimallien käyttämistä OCR:n siivouksessa
- Semantiikkaan ja tekoälyyn perustuva sisällönkuvailu
 - Asennettu skosmos (finto.fi), kun OCR ok sanoille merkityksiä
 - Testattu erilaisia olemassa olevia tekoälyratkaisuita sisällönkuvailuun
 - Mm. tiivistelmien teko
- Fokusryhmä
 - Selvitetty arkistopalveluita tarjoavien tahojen todellisia tarpeita tekoälyyn liittyen
 - Aloitettu Memoriaalin, Dalain ja tämän hankkeen tuloksien yhdistämisen suunnittelu
 - Määrittelytyö





Semantiikkaan ja tekoälyyn perustuva kuvailu

- Asennettu ja käytettävissä
<https://skosmos.memorylab.fi/fi/>
 - Voi käyttää devaamiseen, myös API
<https://skosmos.memorylab.fi/rest/v1/yso/lookup?label=kissa>
 - Päälle rakennettu demo:
<https://memorylab.fi/AIDA/skosmos-demo/>
 - Ei missään nimessä tuotantoon, vaan lähtökohta jatkokehitykseen
- Pienimuotoisia kokeiluita LLM:n käytöstä rakenteiden tunnistamisessa
 - Asennettu (suljettu) OpenChat instanssi millä kokeiltu suljetulla aineistolla
 - LLM:n käyttö OCR:n laadun parantamisessa

Sisältö ja aikataulu

- 12.30 - 12.35 Tilaisuuden avaus, Xamk / Anssi Jääskeläinen
- 12.35 - 12:45 Avauspuheenvuoro, Musiikkiarkisto / Juha Henriksson
- 12.45 - 12.55 [Tekoälyllä lisäarvoa digiarkistojen asiakkaille hanke](#), Xamk / Anssi Jääskeläinen
- **12.55 - 13.55 OCR:n tehostaminen kokonaisuutena**
 - 12.55 - 13.05 Mihin annotointia tarvitaan tekoälyaikana, Elka / Satu Soivanen
 - 13.05 - 13.10 Miten annotointi toteutettiin, Kansallisarkisto / Sini Rajaniemi
 - 13.10 - 13.25 Paddle OCR moottorin jatkokouluttaminen annotoiduilla ja synteettisellä aineistolla, Kansallisarkisto / Atte Föhr & Xamk / Tuomo Räisänen
 - 13.25 - 13.35 Paddle OCR pilotit Xamk / Anssi Jääskeläinen
 - 13.35 - 13.45 Testaajan kokemuksia OCR pilotin käytöstä, Merkki / Johanna Mieto
- 13.45 – 14.10 Kahvitauko & Jatkokehitysideoita yhdessä keskustellen, Hanketiimi
 - Etäosallistujat voivat heitellä ideoita chattiin
- 14.10 - 14.30 Fokusryhmähaastattelujen raportin yhteenveto ja refleктоiva keskustelu, Xamk / Mira Kolari & Musiikkiarkisto / Juha Henriksson
- 14.30 - 14.40 Memoriaalin, Arkkiivin ja tämän hankkeen tuloksien yhdistäminen ja hankejatkumo, Xamk / Mira Kolari
- 14.40 - 14.55 Koneoppimista ja RAG hyödyntämistä, Kansallisarkisto / Mikko Lipsanen
- 14.55 - 15.00 Yhteenveto ja päätös, Xamk / Noora Talsi



**Euroopan unionin
osarahoittama**

Paddle OCR pilotit

Loppukäyttäjät(kö) edellä

Tiedostot

- Inference.pdiparams
- Inference.pdiparams.info
- Inference.pdmodel

→ Niin mitähän näillä pitäisi tehdä..?

Ottaa käyttöön jotakuinkin näin

```
model_path = './models'  
device = picker.getCPUorGPU()  
if (device == "cuda"):  
    ocr = PaddleOCR(lang='latin', det=True, use_angle_cls=True,  
                    rec_model_dir=model_path, show_log=False, use_space_char=True,  
                    use_gpu=True, use_tensorrt=True)  
else:  
    ocr = PaddleOCR(lang='latin', det=True, use_angle_cls=True,  
                    rec_model_dir=model_path, show_log=True, use_space_char=True,  
                    use_tensorrt=True)
```

→ Ei kovin käyttäjäystävällistä ja helppoa

Demot dockeroitu

```
memorylab-aj@CPU-portaine x + v
memorylab-aj@CPU-portainer:~/dockers/extendedPaddleOCR$ ls -al
total 28
drwxrwxr-x 4 memorylab-aj memorylab-aj 4096 May 16 06:52 .
drwxrwxr-x 7 memorylab-aj memorylab-aj 4096 Aug 12 06:29 ..
-rw-rw-r-- 1 memorylab-aj memorylab-aj 377 May 16 06:54 Dockerfile
-rw-rw-r-- 1 memorylab-aj memorylab-aj 5439 May 21 09:53 extendedPaddleOCR.py
drwxrwxr-x 2 memorylab-aj memorylab-aj 4096 May 17 05:02 models
drwxrwxr-x 2 memorylab-aj memorylab-aj 4096 May 13 06:59 snakeoil
memorylab-aj@CPU-portainer:~/dockers/extendedPaddleOCR$ |
```

- 2 komentoa
 - `sudo docker build -t extendedpaddleocr .`
 - `sudo docker run -p 8087:8087 --name extendedpaddle -d --restart unless-stopped extendedpaddleocr`
- Näiden jälkeen selaimella ip-osoite/domain:8087
 - Esim: <https://192.168.10.108:8087/>
- Julkidemo: <https://memorylab.fi/AIDA/extended-paddle-demo/>
 - Apache .conf tiedostossa: ProxyPass /AIDA/extended-paddle-demo/ https://192.168.10.108:8087/

Konepellin alla

- Tässä tapauksessa yksi Python tiedosto
 - Koko koodi < 100 riviä
1. Otetaan vastaan ladattu tiedosto
 - Mahdollinen tiff → jpg muunnos
 2. Luetaan kuva jollakin Python kirjastolla (pil, pillow, yms.)
 3. Syötetään luettu kuva alustetulle Paddle OCR moottorille
 4. Otetaan ocr tulokset vastaan
 - Piirretään laatikot
 - Lasketaan tarkkuus
 - Tallennetaan tekstinä

Uppaa tähän kuva
ruotsi-sample.jpg 244.9 KB

OCR-lue

Tulokset

OCR tulokset

Räntevinst
Ford-Touring. Pris kr. 1875 fob. Malmö inkl. gummiskatt, plus järnvägstrakt till köparens ort. Med självstart 100 kr. extra
Skall Ni välja en vagn?
Tjänstemäns och
Varje fabrik reklamerar med harelivväl icke Ford nämnda
Direktionsarvode
Revisionsarvode
en bestämd egenskap hos sin egenskaper? Fråga en Ford
Hyreskostnad
vagn: Den är stark eller den ägare, eller lägg blott märke
Skatter
är bekväm och pålitlig eller till hur många Fordvagnar
Pension
den är praktisk
Ni möter på ga
Diverse omkostna
1875

Tarkkuus

94.05

Lataa tekstinä
ruotsi-sample.jpg.txt 820.0 B

Huom! Tarkkuus perustuu ainoastaan PaddleOCR:n omaan riviakohtaiseen arvioon tunnituksen tarkkuudesta.
Kuvassa punainen < 75%, oranssi 75-85%, keltainen 85-95% ja vihreä yli 95%

Laatikoitu kuva

Use via API - Built with Cradio

Käyttöliittymät on vaikeita rakentaa..

- Mutta kun ei enää ole 😊.
- Käyttöliittymä tehty Gradiolla → Todella helppo
- Tässä 1 rivi ja käynnistys
- Demossa ~20 riviä

```
import gradio as gr

def greet(name):
    return "Hello " + name + "!"

demo = gr.Interface(fn=greet, inputs="text", outputs="text")
demo.launch()
```



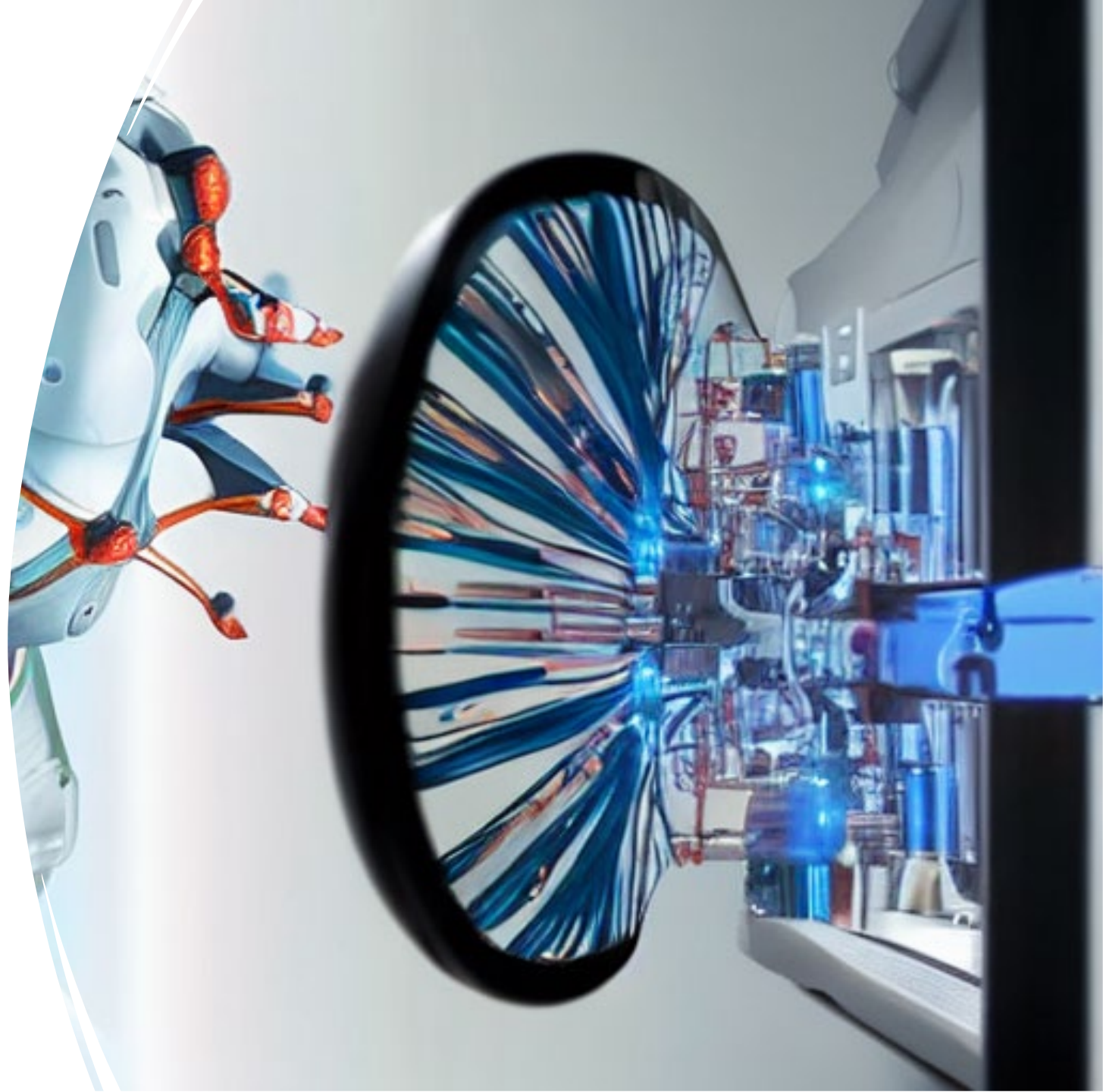
The screenshot shows a web interface for a simple chatbot. On the left, there is a text input field labeled "name" with a small "x" icon in the bottom right corner. Below the input field are two buttons: a grey "Clear" button and an orange "Submit" button. On the right, there is a text output field labeled "output" with a small "x" icon in the bottom right corner.

Toinen demo

- Sama taustalogiikka mutta toimii .pdf tiedostoille ja näyttää olemassa olevan tekstidatan jos sellainen on.
- <https://memorylab.fi/AIDA/PDFocr-with-boxes/>
 - HKA:lta tuli jo toive että ocr tieto mukaan .pdf tiedostoon
 - Tehtävissä, ei hankkeen puitteissa ajallisesti mahdollinen
- Molempien demojen koodit (ja muutakin) saatavissa xamkfi githubin kautta
 - <https://github.com/orgs/xamkfi/repositories>
 - digitalia-aida alkuiset
 - <https://github.com/xamkfi/digitalia-aida-extended-paddle-demo>
 - <https://github.com/xamkfi/digitalia-aida-pdf-paddleocr>

Yhteenveto

- Pelkkä uusi tekniikka ei helpota käyttöä
- Tavalliset käyttäjät eivät osaa/halua/suostu käyttämään komentoriviä
 - Tarvitaan helppokäyttöinen UI jolla asiaa voidaan testata ilman teknistä osaamista





Tauko

- Jatketaan 14.10
- Etäosallistujat voivat kirjoitella kehitystarpeita chatiin tai
- Koko tilaisuuden ajan käytössä
 - <https://www.menti.com/aldowixac9jy>
 - Palautetta, kommentteja, kehitysehdotuksia, jne.