

Memory Lab – tekoälyä digiarkistoihin ja arkistojen digiin

TkT. Anssi Jääskeläinen

Tutkimuspäällikkö

Xamk / Digitaalinen

XAMK

Agenda

Memory Lab: Mitä, Miksi,
Milloin?

Tekoälyä digiarkistoihin,
esimerkkien kautta

Mitä hyötyä Memory
Labista on loppukäyttäjille



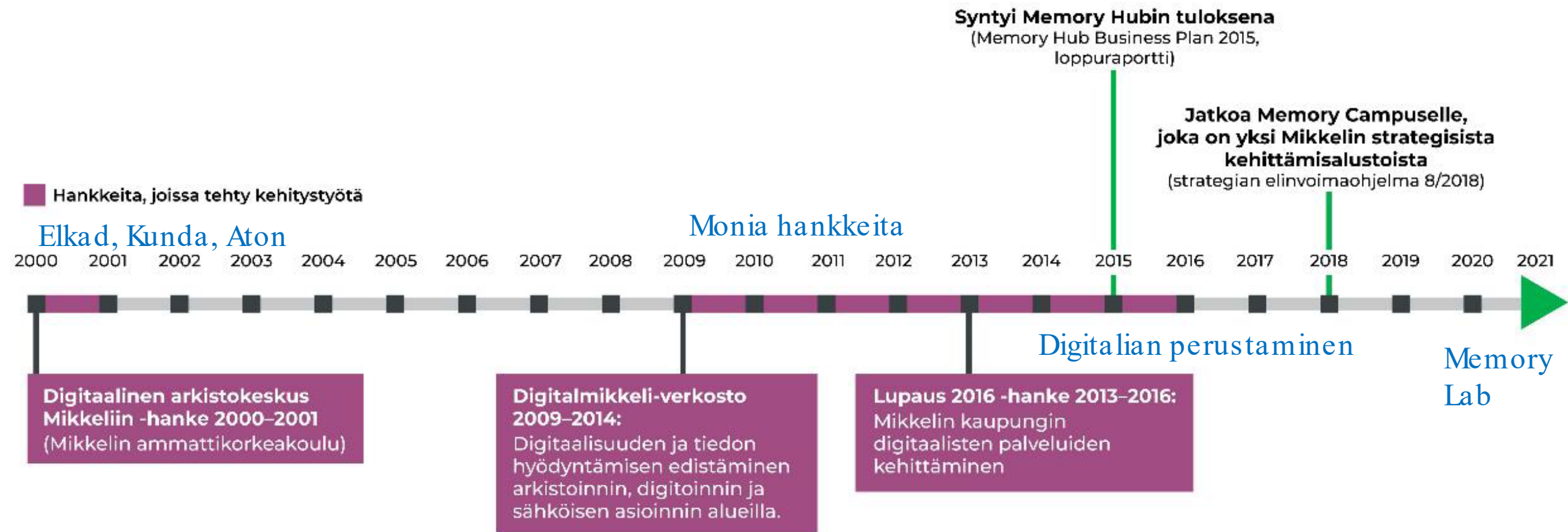
Toista iseksi - rautaa

- Teorettinen huippulaskentateho on 5 petaflopsia (6,5kW)
- Laskentateho vastaa noin 20 000 peruskannettavan yhteenlaskettua tehoa (vrt. LUMI 1 500 000).
- Nvidian A100 GPU kortteja
- “Perus palvelin kultaisilla kuorilla”
- Lisätietoja memorylab.fi



Tulevaisuudessa, mahdollisesti..

Memory Lab taustat



Tekoälyä digiarkistoihin

- Laadunvalvonta arkistoon viedessä → **Tarkastukset**
- OCR:n laatu vanhalla materiaalilla on usein huono → **Laadun parannus**
- Kuvailutietojen syöttäminen on tylsää → **Automaattiset kuvailutiedot**
- Vanhat tekniikat eivät tunnista kuvien sisältöjä → **Erilaisia tekoälypohjaisia tunnistusratkaisuja**
- Isojen aineistomassojen annotointi vie aikaa → **Synteettinen data**
- Miten tietää nopeasti mistä isossa dokumentissa kerrotaan → **Automaattinen tiivistelmien teko**

OCR laadun parannus

Malli / CER	Elka testi (4273 kuvaa)	DALAI testi (3475 kuvaa)	KA testi (3247 kuvaa)	ELKA testi käsi (714 kuvaa)
Tesseract	0,046	0,027	0,044	0,793
Alkuperäinen	0,067	0,039	0,068	0,502
Koulutettu	0,020	0,012	0,023	0,207

Taulukko: Lisäkoulutuksella saatu hyöty: Atte Föhr, Kansallisarkisto

1. Koulutetaan avointa OCR moottoria omalla aineistolla

- Annotoitu / synteettinen
 - Koulutettumalli: https://github.com/project-AIDA/Finnish_PaddleOCR/tree/main
 - Treenidata: https://huggingface.co/datasets/Kansallisarkisto/AIDA_ocr_training_data
 - Pohjamalli: https://huggingface.co/Kansallisarkisto/PaddleOCR_training
 - Demo (kuvat): <https://memorylab.fi/AIDA/extended-paddle-demo/>
 - Demo (pdf): <https://memorylab.fi/AIDA/PDFocr-with-boxes/>

2. Käytetään OCR tuloksen parantamiseen kielimallia. Testi Copilot

- Käytetään OOR tuloksen parantamiseen kielimallia. → Käytetään OCR-tuloksen parantamiseen kielimallia.
- The maxillipeds, lower edge of the carapace, margins of the sternum and abdomen, and especially the anterior portion of the sternum are fringed with long hair. Legs hairy, except the distal two-thirds of the dactyls.
- → "The maxillipeds, lower edge of the carapace, margins of the sternum and abdomen, and especially the anterior portion of the sternum are fringed with long hair. Legs hairy, except the distal two-thirds of the dactyls."
- It looks like this text is describing the anatomical features of a crustacean or similar organism. If you need any more help, feel free to ask!

Paddle jatkokoulutus	aika
CPU	~250d
Hipun VM (2x A100)	~24h
Hippu (bare metal 8XA100)	~1,5h

Taulukko: Koulutukseen kuluneita aikoja

Esimerkki PaddleOCR vs. Tesseraact

PaddleOCR SE NYT ON YKSI MERKILLINEN ASIA, ETTÄ MEIDAN EI SALLITA PUKEA EDES KOULULAISIA MUSTIIN PAITOIHIN, VAIKKA NI KIRJOITTA - , ETTÄ KUN SUOMEN TALONPOJAT TULIVAT REP= PUINEEN MÄNTSÄLÄÄN SYÖMÄÄN, NIIN HALLITUS TULI TYKKIEN KANSSA HEIDÄN RUOKARAUHAANSA HÄIRITSEMÄÄN. AVUN KAIKEN TÄLLÄISEN ESTÄMISEKSI ON LÄHDETTÄVÄ KANSAN RIVEISTÄ. RIIPNNEN

Tesseraact SE NYT ON YKSI MEEKILLINEN ASIA/ETTÄ, MEIDAN Ez .SALLITA Pust spss KOULULAISIA MUST/IN FAITOIHIN mmm VASEMMALLA 75 KANSANEDUSTAJAA SAA VALMISTAA KAPINAA —" JA OLI SE NYT SITTEN u-SEKINJUTTU KUTEN VIROLÄ'NEN YSTAVA: NI KIBJOITTA—V ETTÄ, KUN" SUOMEN TALONPQJAT TULIVAT REP= FWNEEN MÄNTSÄLÄÄ'N SYOMAAN, NIIN HALLITUS TULI TYKKIEN KANSSA HEIDAN RUOKARAUHAANSA HAIRITSEMÄANA. - AVUN KAIKEN TALLÄISEN ESTAMISEKSI ON LAHDETTAVÄ KANSAN RIVEISTA" .



SE NYT ON YKSI MERKILLINEN ASIA, ETTÄ MEIDÄN EI SALLITA PUKEA EDES KOULULAISIA MUSTIIN PAITOIHIN, VAIKKA VASEMMALLA 78 KANSANEDUSTAJAA SAA VALMISTAA KAPINAA. — JA OLI SE NYT SITTEEN SEKIN JUTTU — KUTEN VIROLAINEN YSTÄVÄ — NI KIRJOITTA —, ETTÄ KUN SUOMEN TALONPOJAT TULIVAT REP=PUINEEN MÄNTSÄLÄÄN SYÖMÄÄN, NIIN HALLITUS TULI TYKKIEN KANSSA HEIDÄN RUOKARAUHAANSA HÄIRITSEMÄÄN. — AVUN KAIKEN TÄLLÄISEN ESTÄMISEKSI ON LÄHDETTÄVÄ KANSAN RIVEISTÄ.

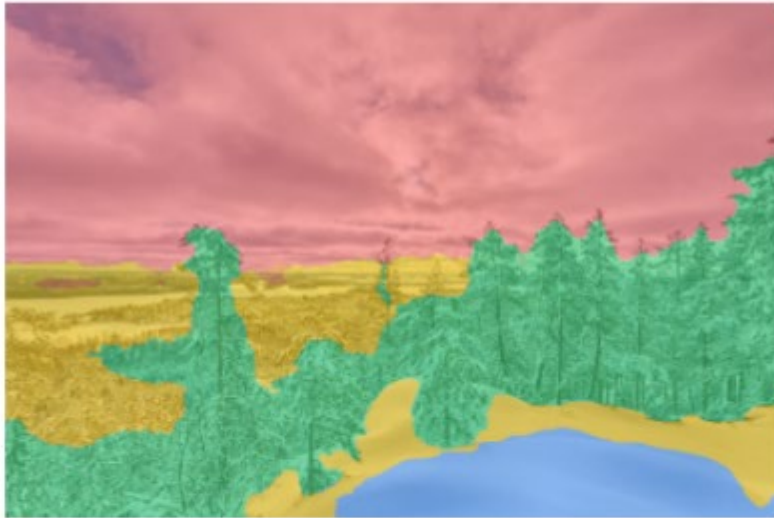


Kuvailutietojen syöttö – ei kivaa

- Työkseen tekevät / kansalaiset
- Vois ikohan tätäkin helpottaa jotenkin → Arkkiivi, <https://arkkiivi.fi/>
- <https://annif.org/>
- Kuvalle automaattiset asiasanat: historia, ruokakulttuuri, ruoka-aineet, ensimmäinen maailmansota, koululaiset, ruokaohjeet, puut (kasvit), virolaiset, sotataito, ravintoarvo

Voisiko myös kuvista tunnistaa asioita?

- <https://huggingface.co/facebook/detr-resnet-50>
- <https://huggingface.co/nvidia/segformer-b0-finetuned-ade-512-512>



sky	1.000
tree	1.000
earth	1.000
mountain	1.000
land	1.000



person	0.865
person	0.844

Synteettinen data

- Ja tämä tarkoittaa mitä?
- Edut:
 - Säästää huomattavasti aikaa annotointiin nähden
 - 100% oikein

mä olivat poliittisesti turvallisia aihe-

kallistui niin syvään, että sen äärimmäiset oksat hipoivat kirkasta

Perjantai-aamuna tulivat suureen sydänmaan kylään, jossa matka oli

"Kyllä", sanoi Ebba lujasti, "minä luulen, että sinä olet tehnyt

Kuvat yllä: Atte Föhr, Kansallisarkisto



Automaattiset tiivistelmät

- Valtakielillä erinomaiset tulokset avoimillakin malleilla
 - <https://huggingface.co/facebook/bart-large-cnn>
- Suurten toimijoiden mallit Suomellakin vähintäänkin kohtuullisesti
- Testi Copilot:
 - Lähtötieto: Teksti sivulta: <https://www.xamk.fi/tapahtuma/digital-archiving-futures-digiarkistojen-tulevaisuuskuva/>
 - Tulos: Memory Campus ja Digitalia järjestävät kolmipäiväisen hybriditapahtuman “Digital Archiving Futures – Data Spaces, Artificial Intelligence, and Cluster Collaboration” 4.–6. syyskuuta 2024. Tapahtuma alkaa yhteistyökumppaneiden kokoontumisilla ja avoimella iltatapahtumalla Sodan ja rauhan keskus Muistissa. Toisena päivänä kansainväliset asiantuntijat keskustelevat englanniksi data-avaruuksista, tekoälystä ja klusteriyhteistyöstä. Kolmantena päivänä Digitalian perinteinen Digitaalisen tiedon kesäkoulu kokoaa kotimaisia puhujia ja yleisöä jatkamaan keskustelua digiarkistoinnin tulevaisuudesta. Lisätietoja ja ilmoittautuminen: Kati Saltiola, kati.saltiola@xamk.fi.

Yhteenveto

- Tekoäly on tullut jäädäkseen
- Ihmisen oltava loopissa mukana
 - Yksinkertaisimmat asiat hoitunevat täysin automaattisesti
- Huggingface == Tekoälyn github (~900 000 avointa mallia)
- Tekoälyä voi ajaa myös CPU:lla
 - Kouluttaminen vaatii tehoja
- Memory Lab & Digitalia edistää tekoälyn hyödyntämistä kokeilujen avulla

