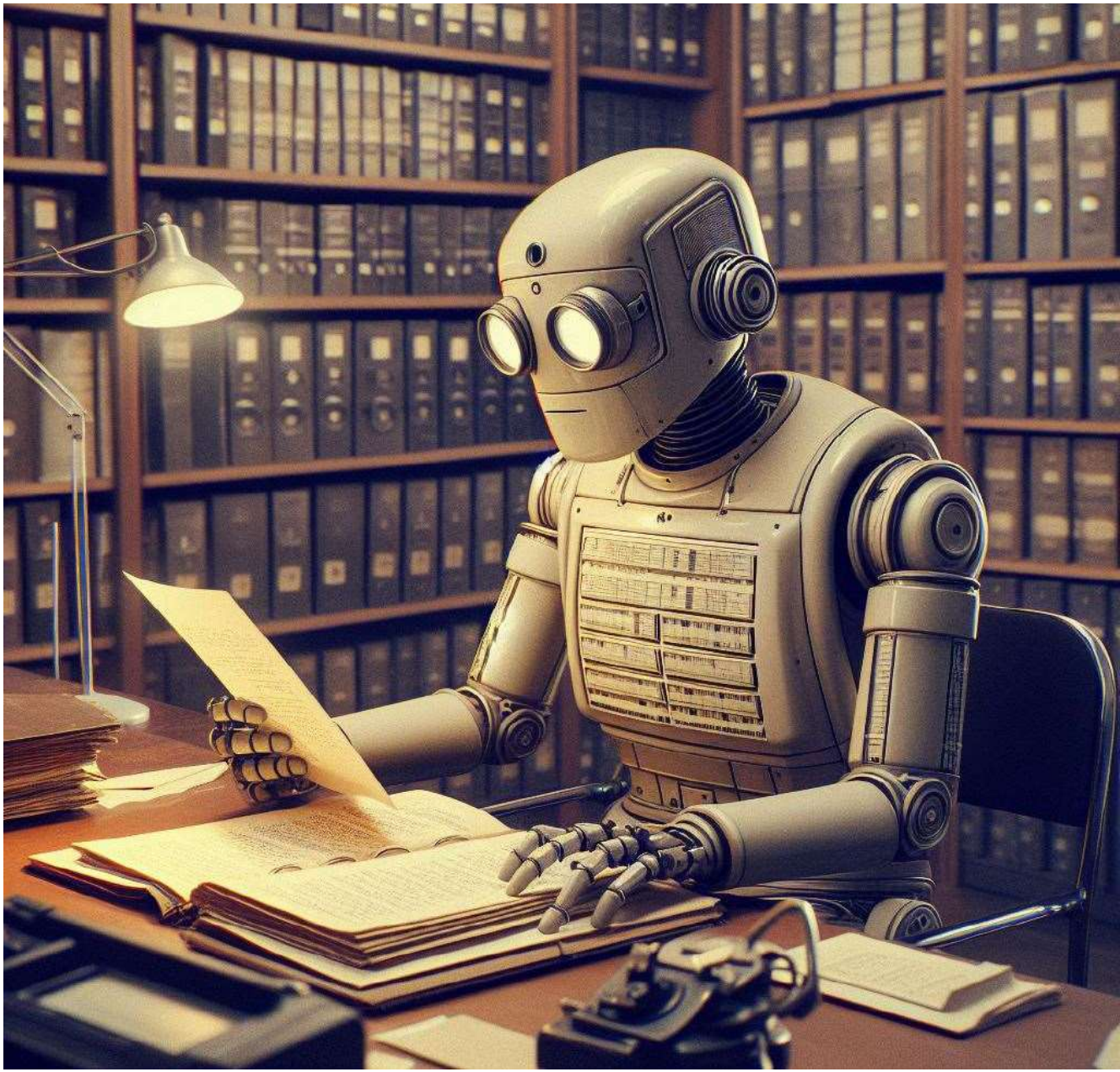
KANSALLISARKISTO

Tekoäly asiakirjahallinnassa

Ilkka Jokipii, yksikönpäällikkö

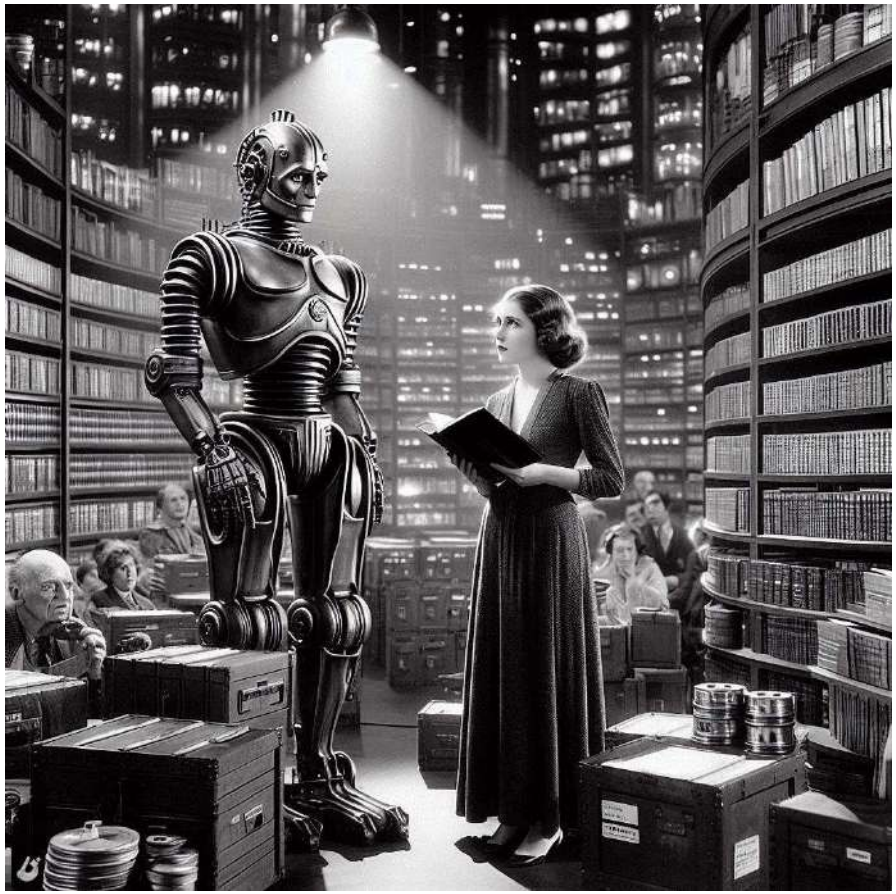
Mikko Lipsanen, koneoppimisen pääsuunnittelija

13.4.2024



Luennon sisältö:

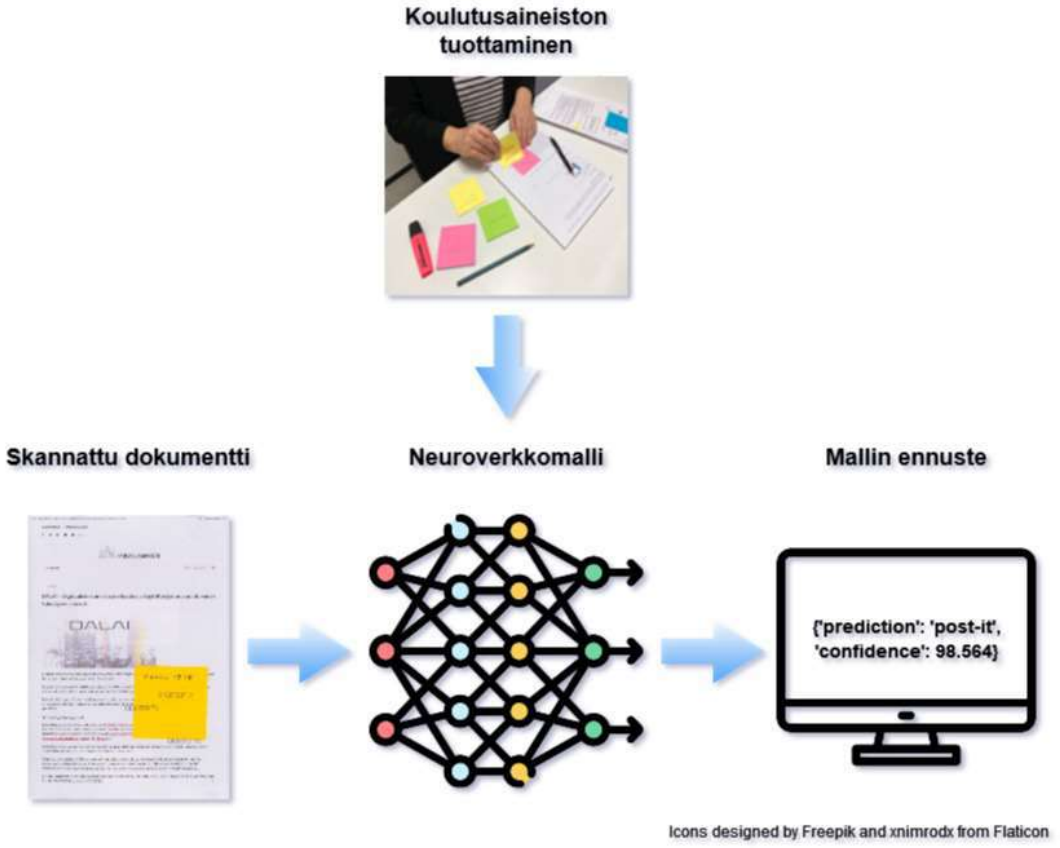
1. Lyhyesti tekoälystä
2. Tekoälyn hyödyntäminen asiakirjahallinnan eri vaiheissa
   - Asiakirjan digitointi
   - Digitoidun asiakirjan laadun validointi
   - Asiakirjan sisällön tunnistus
   - Tunnistetun (teksti)sisällön analysointi
3. Käytettävyys: miten tekoälysovelluksiin pääsee käsiksi?
4. Tekoäly Kansallisarkistossa



KANSALLISARKISTO

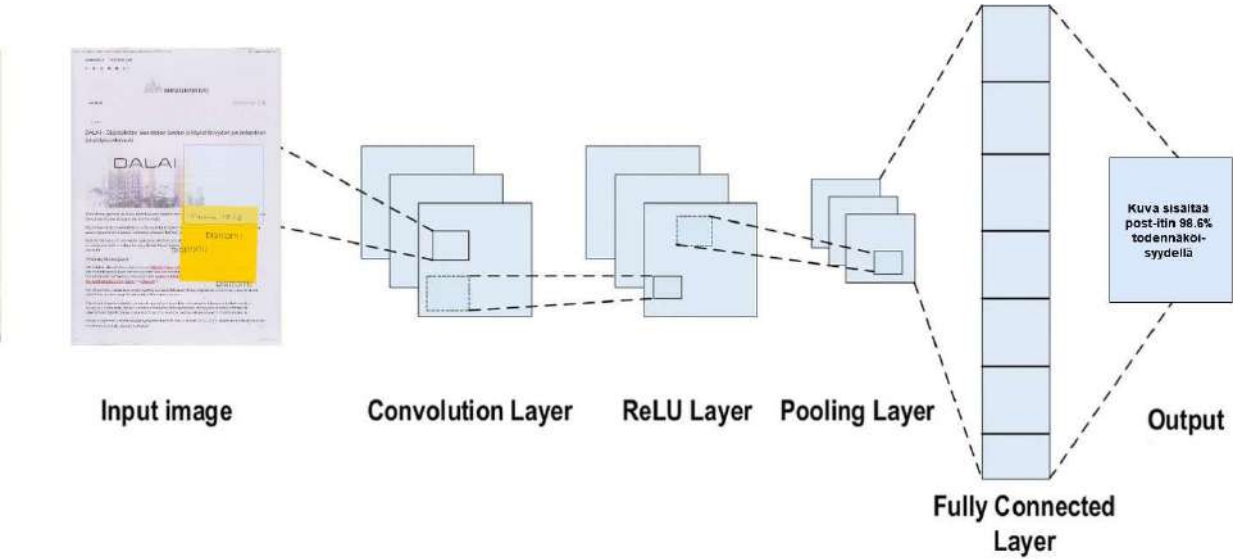
1. Lyhyesti tekoälystä

- **Tekoäly**: tuo helposti mieleen ihmisen kaltaisen älykkään toimijuuden, vaikka taustalla on yleensä paljon pienimuotoisempia, erilaisiin tehtäviin erikoistuneita koneoppivia sovelluksia
- **Koneoppiminen**: hyödyntää algoritmeja joiden avulla mallit oppivat aineistosta asioita joita niihin ei ole ennalta 'ohjelmoitu': kyky yleistää annetuista lähtökohdista
  - ***Ohjattu koneoppiminen***: sovellus oppii etsimään aineistosta haluttuja asioita sille annetun rajallisen esimerkkiaineiston perusteella
  - ***Ohjaamaton koneoppiminen***: aineistoa luokitellaan sen piirteiden ja säännönmukaisuuksien perusteella



KANSALLISARKISTO

- **Syväoppivat neuroverkkomallit** dominoineet koneoppimisen eri sovelluskohteita (kuvien tunnistus, luonnollisen kielen prosessointi (NLP), robotiikka..) viimeistään 2010-luvun lopulta alkaen
  - Tulokset usein ylivoimaisia aiempiin tekniikoihin nähden
  - Haasteena kuitenkin mallien vaikea tulkittavuus (’musta laatikko’): päättelyr logiikkaa ja virheiden syitä usein vaikea selvittää, kontrollointi ja tulosten tarkkailu tärkeää



Alkuperäinen kuva: Alzubaidi, L., Zhang, J., Humaidi, A.J. *et al.* Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* 8, 53 (2021).

- **Suurten kielimallien** kehitys tuo mukanaan paljon uusia mahdollisuuksia, mutta niiden hyödyntäminen esim. arkistosektorilla on vielä alkuvaiheessa
  - Voivat helpottaa ja nopeuttaa prosesseja joihin aiemmin tarvittu useita erikseen koulutettuja malleja
  - Toisaalta voivat ’keksiä’ omia sisältöjään, kontrollointi usein haastavaa
  - Kaupallisten toimijoiden ja englanninkielisten mallien dominanssi
    - Kehitteillä kuitenkin uusia suomenkielisiä avoimia kielimalleja (TurkuNLP, SiloAI)

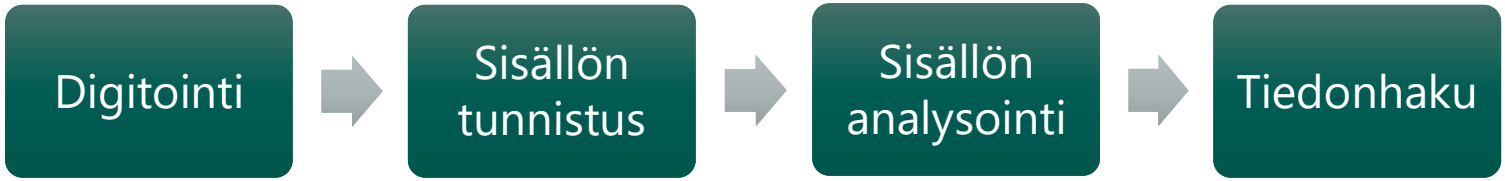
Esimerkki neuroverkkomallien haasteista: tekstintunnistusmallin tulkinta taulukkosolun sisällöstä.

Miksi hyödyntää tekoälyä asiakirjahallinnassa?

Tulokset	• Neuroverkkojen kehityksen myötä tekoälymallien nopeus ja tulosten laatu tehneet niistä aiempaa kilpailukykyisempiä.
Yleistyvyys	• Manuaalisesti määriteltyihin sääntöihin perustuvia malleja on usein vaikea sovittaa aineiston moninaisuuteen, kun taas koneoppivat mallit pystyvät tähän paremmin.
Käytettävyys	• Koneoppimisen hyödyntäminen on tullut helpommaksi: valmiit sovellukset, esikoulutetut mallit, uudet ohjelmointikirjastot, pilvipalvelut jne.

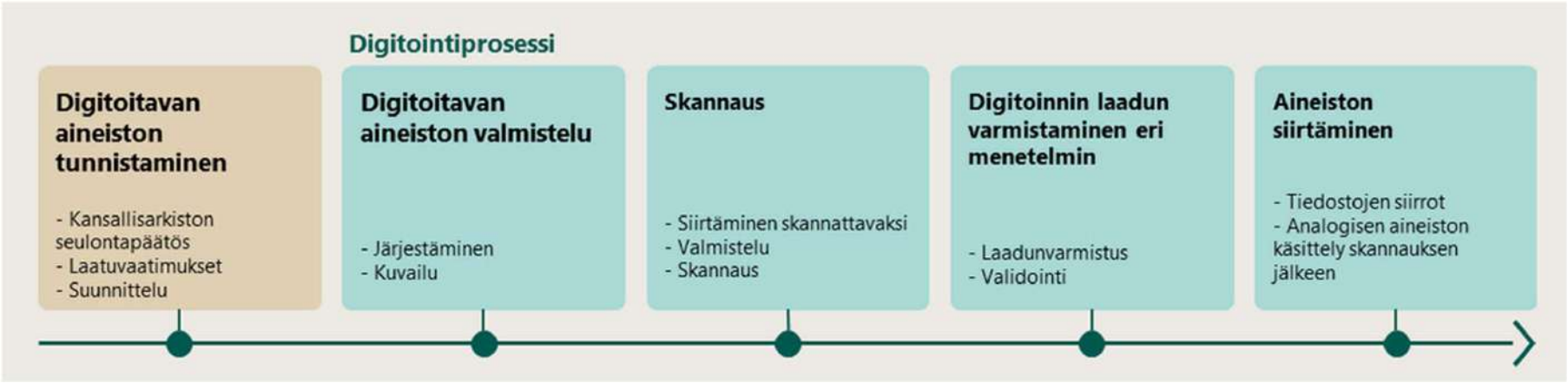
2. Tekoäly asiakirjahallinnan eri vaiheissa

- Käydään läpi tekoälyn hyödyntämisen mahdollisuuksia erityisesti arkistoaineiston näkökulmasta
- Runsaasti käytännön esimerkkejä!



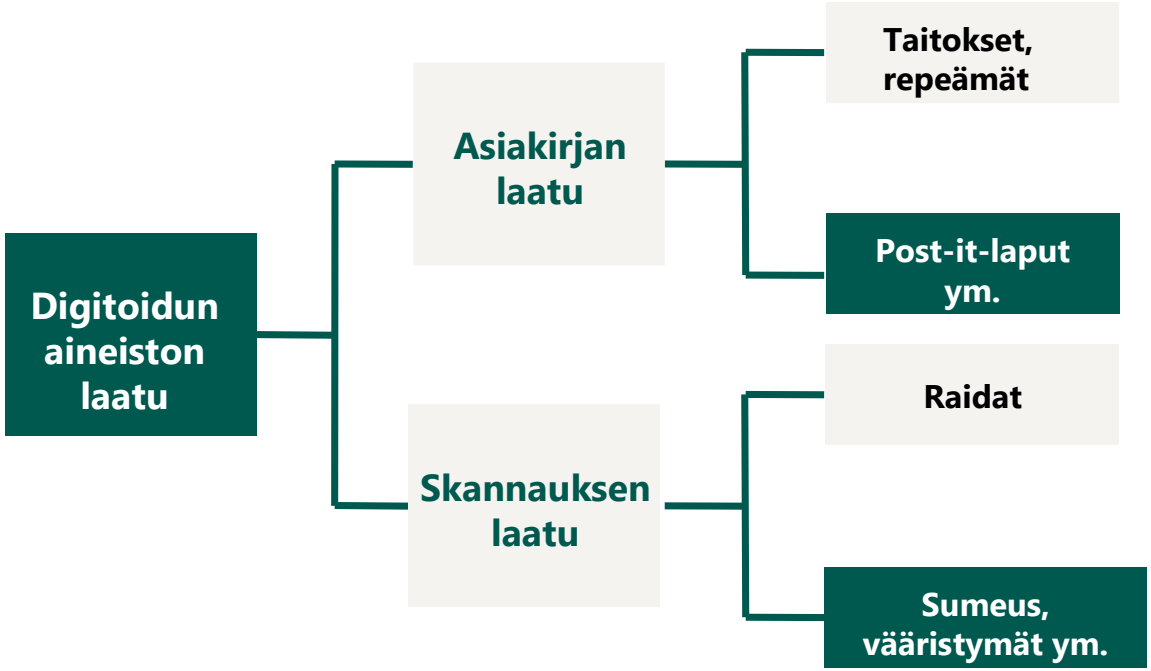
Asiakirjan digitointi

- Digitointiprosessin vaiheet Kansallisarkistossa (https://kansallisarkisto.fi/digitointi)



- Tekoälysovelluksilla voi olla rooli jo skannausvaiheessa, mikäli skannerin ohjelmisto hyödyntää sellaisia esim. kuvan laadun tarkkailuun ja optimointiin
- Keskitytään kuitenkin neljänteen vaiheeseen: digitoinnin laadun varmistus ja validointi

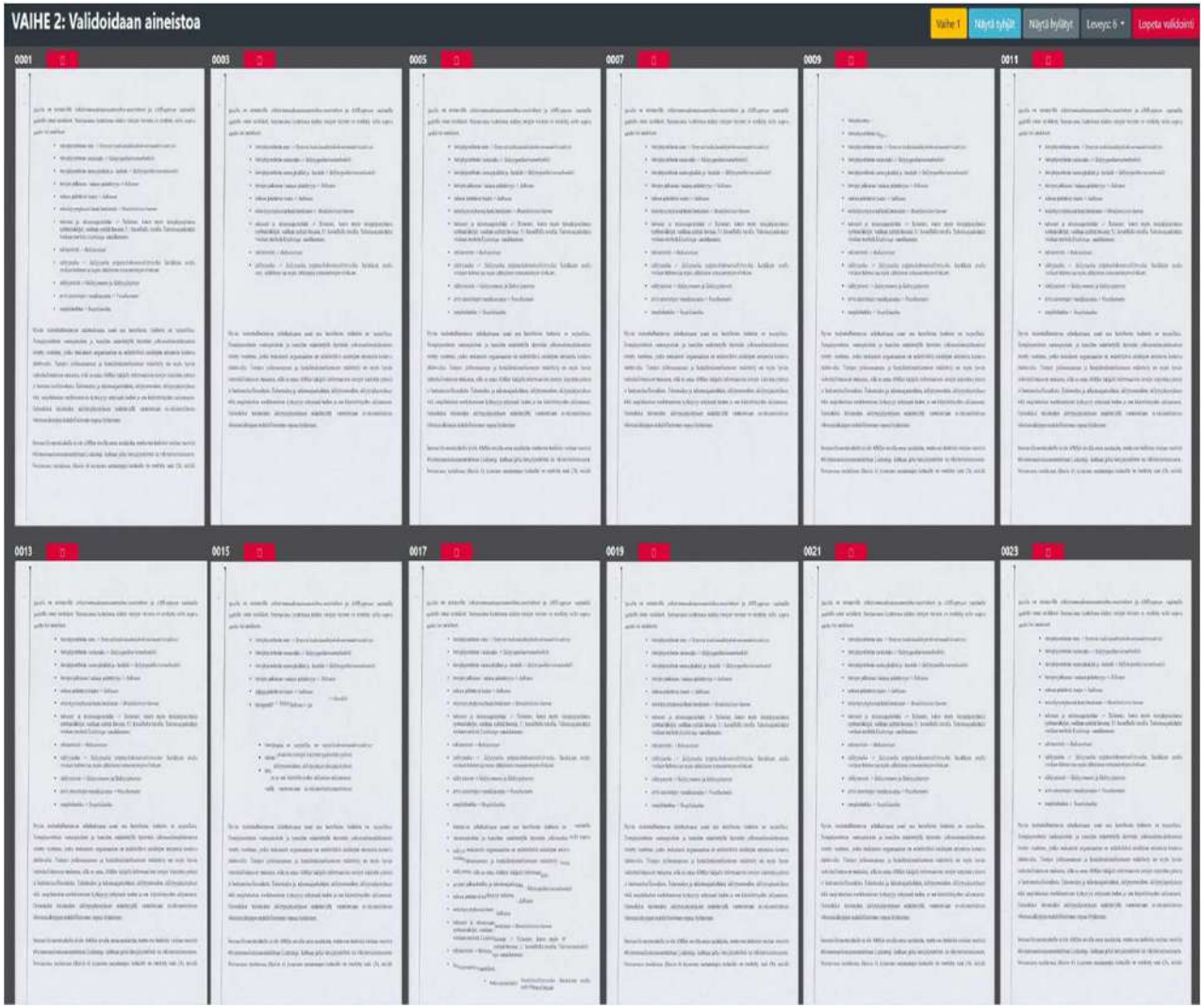
Digitoidun asiakirjan laadun validointi



Automatisoinnin hyödyt

- Manuaalinen laadun tarkkailu vaatii paljon resursseja, kun aineistomäärät ovat suuria
  - Esimerkiksi kansallisarkiston massadigitoinnissa* vuosittainen digitointitavoite on 1800 hyllymetriä!
- Automatisoinnin avulla voidaan säästää aikaa ja siten myös rahaa - edellyttäen että käytetyt työkalut toimivat riittävän hyvin

**Massadigitointi: valtion viranomaisten arkistoitavien asiakirjojen digitointi pysyvää säilytystä varten.*



Kansallisarkiston massadigitoinnissa käytetty digitoidun aineiston manuaalisen validoinnin käyttöliittymä.

KANSALLISARKISTO

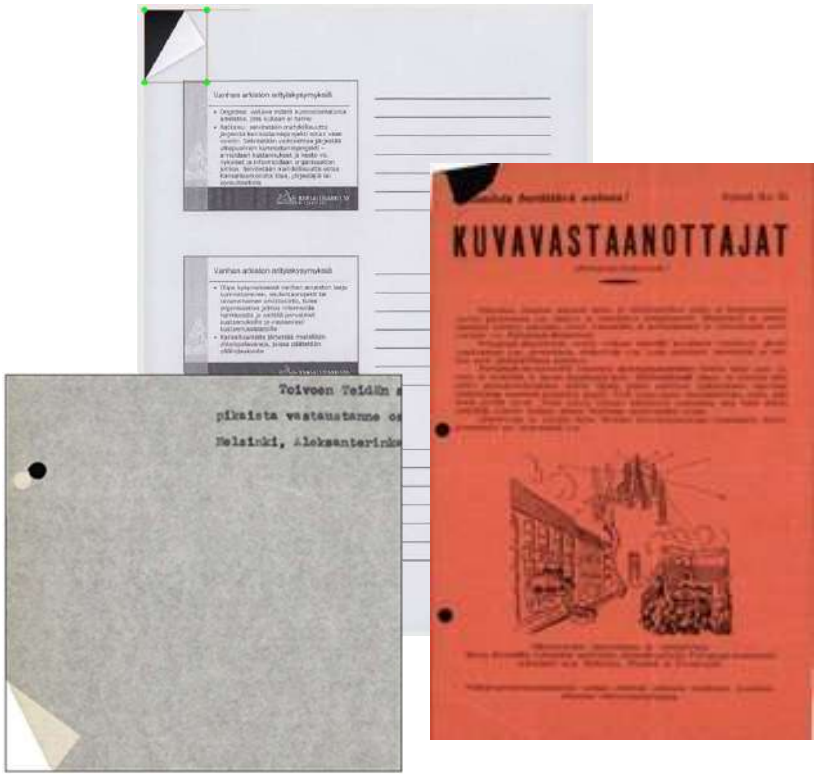
Automatisoinnin haasteet

- Aineiston ja tunnistettavien elementtien moninaisuus on haaste työkalujen yleistettävyydelle
- Koneoppivien mallien koulutukseen tarvitaan oikeanlaista aineistoa ja usein myös manuaalista annotointia*

**Annotointi: aineiston luokittelua / kuvaamista koulutettavan mallin kannalta relevantilla tavalla.*

Taittuneiden kulmien tunnistus

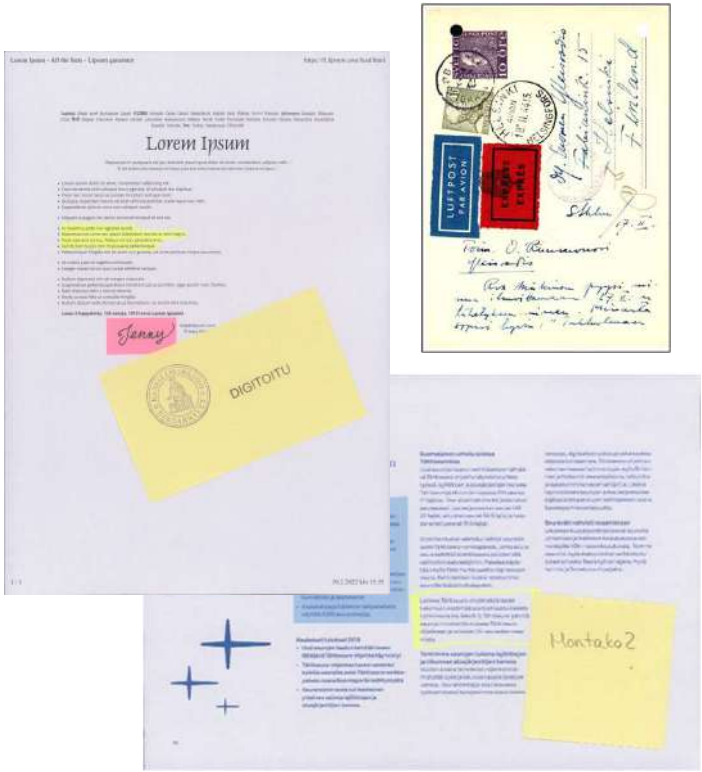
- Sovellus saa syötteenä dokumenttisivun sisältäviä kuvatiedostoja ja luokittelee ne kahteen luokkaan: sivu sisältää taittuneen kulman / sivu ei sisällä taittunutta kulmaa
- Tunnistaa myös repeämät paperin reunoissa
- Neuroverkkomallin koulutusta varten koottiin aineisto jossa oli tuhansia eri tavoin taittuneita ja repeytyneitä asiakirjoja, sekä vastaavasti erilaisia ehjiä asiakirjoja



KANSALLISARKISTO

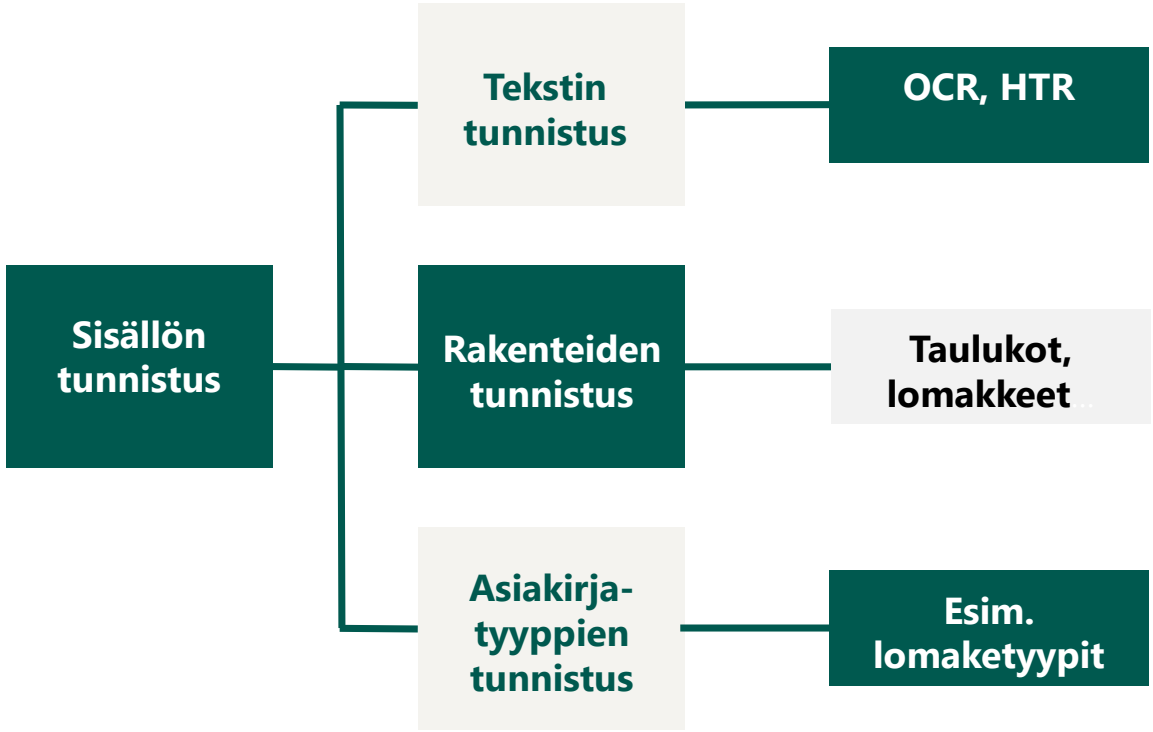
Post-it-lappujen tunnistus

- Digitoija ei aina huomaa post-it-lappuja, jotka voivat peittää dokumentin sisältöä
- Kun post-it huomataan, Kansallisarkiston massadigitoinnissa dokumentti digitoidaan sekä lapun kanssa että ilman sitä
- Sovellus luokittelee sivut kahteen luokkaan: sivu sisältää post-it-lapun / sivu ei sisällä post-it-lappua
- Neuroverkkomallin koulutusaineistoa askarreltiin paljon käsin: eri värisiä ja muotoisia post-it-lappuja tekstillä tai ilman erilaisiin dokumentteihin liimattuina



KANSALLISARKISTO

Digitoidun asiakirjan sisällön tunnistus



Automatisoinnin hyödyt

- Etsityn tiedon sekä aineistojen välisten yhteyksien löytäminen on vaikeaa jos toimitaan vain digitoitujen kuvien tasolla
- Tekstisisällön manuaalinen transkribointi on hidasta: koneoppivat sovellukset nopeuttavat prosessia

Automatisoinnin haasteet

- Tekstisisällön tunnistuksen laatu vaihtelee, vaikuttaa metatietojen tunnistuksen laatuun
- Rakenteelliset elementit kuten taulukot sisältävät valtavasti variaatiota: tunnistuksessa usein toisessa vaakakupissa laatu, toisessa yleistettävyys
  - Tulokset yleensä parempia kun mallit räätälöidään tietyn tyyppiselle aineistolle

Tyhjien sivujen tunnistus

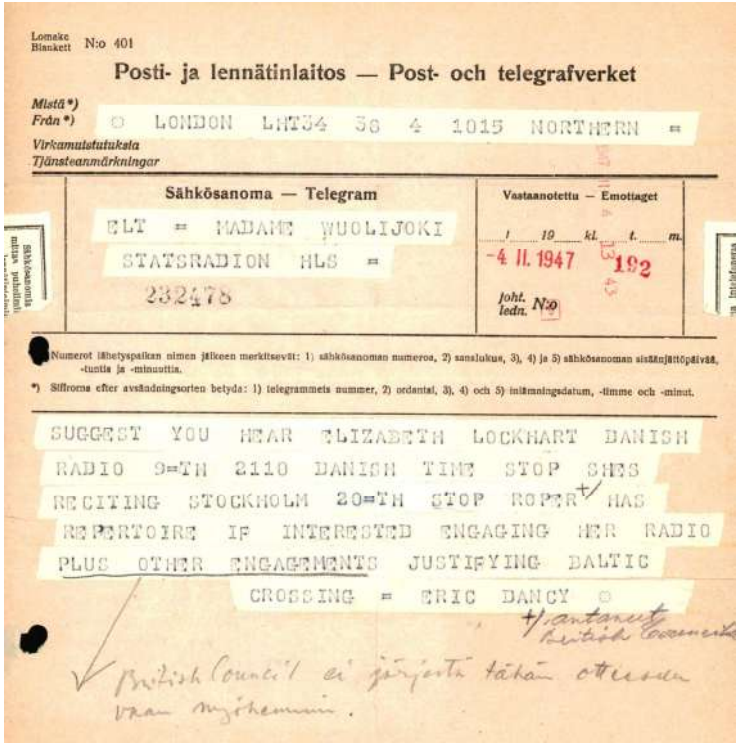
- Lähes 40% Kansallisarkiston massadigitoinnissa digitoiduista sivuista on tyhjiä
- Tyhjiäkään sivuja ei tuhota, mutta niitä ei haluta ajaa sisällöntunnistuksen läpi tai viedä käyttöliittymään
    - Siksi ennen sisällön tunnistusta on tärkeää erotella tyhjät sisällöllisistä dokumenteista
- Tyhjien tunnistusta varten koulutettiin koneoppimismalli, koulutusaineistossa yli 100 000 kuvatiedostoa
- Jopa tyhjän sivun tunnistus voi olla yllättävän haastava tehtävä: esimerkiksi sivun läpi heijastuva arkin kääntöpuolen teksti aiheuttaa ongelmia

KANSALLISARKISTO

Konekirjoitetun tekstin tunnistus

- **OCR** (Optical Character Recognition) on keskeinen vaihe digitoidun asiakirjan prosessoinnissa
  - Tekstisisällön muuttaminen digitaaliseen muotoon manuaalisesti erittäin aikaa vievä prosessi
  - OCR:n laatu vaikuttaa kaikkiin prosessoinnin myöhempiin vaiheisiin, joissa hyödynnetään asiakirjan tekstisisältöä
- Tarjolla lukuisia ilmaisia (Tesseract, EasyOCR, PaddleOCR..) ja kaupallisia (Amazon Textract, Microsoft Read OCR, Google Document AI..) sovelluksia
- Mitä rakenteisempi asiakirja on, ja mitä heikompi asiakirjan tai skannauksen laatu, sitä huonolaatuisempi on myös OCR:n tulos
- Tesseract yleisesti (myös Kansallisarkistolla) käytetty ilmaissovellus
- Kansallisarkisto mukana hankkeessa jossa tutkitaan, voidaanko sovellusten laatua parantaa jatkokouluttamalla niitä omalla aineistolla

Konekirjoitetun tekstin tunnistus



Lomake Blankett N:o 401

Posti- ja lennätinlaitos — Post- och telegrafverket

Mistä *) / Från *) O LONDON LHTSJ4 38 4 1015 NORTHERN =

Virkamuistutuksia / Tjänsteanmärkningar

Sähkösanoma — Telegram

ELT = MADAME WUOLIJOKI

STATSRADION HLS =

232478

Vastaanotettu — Emottaget

-4 II. 1947 192

Joht. / ledn. N:o

Numerot lähetyspaikan nimen jälkeen merkitsevät: 1) sähkösanoman numeroa, 2) sanalukua, 3), 4) ja 5) sähkösanoman sisäänjättöpäivää, -tuntia ja -minuuttia.

*) Siffrorna efter avsändningsorten betyda: 1) telegrammets nummer, 2) ordantal, 3), 4) och 5) inlämningsdatum, -timme och -minut.

SUGGEST YOU HEAR ELIZABETH LOCKHART DANISH RADIO 9=TH 2110 DANISH TIME STOP SHES RECITING STOCKHOLM 20=TH STOP ROPER +/ HAS REPERTOIRE IF INTERESTED ENGAGING HER RADIO PLUS OTHER ENGAGEMENTS JUSTIFYING BALTIC CROSSING = ERIC DANCY O

+/ antanut British Council

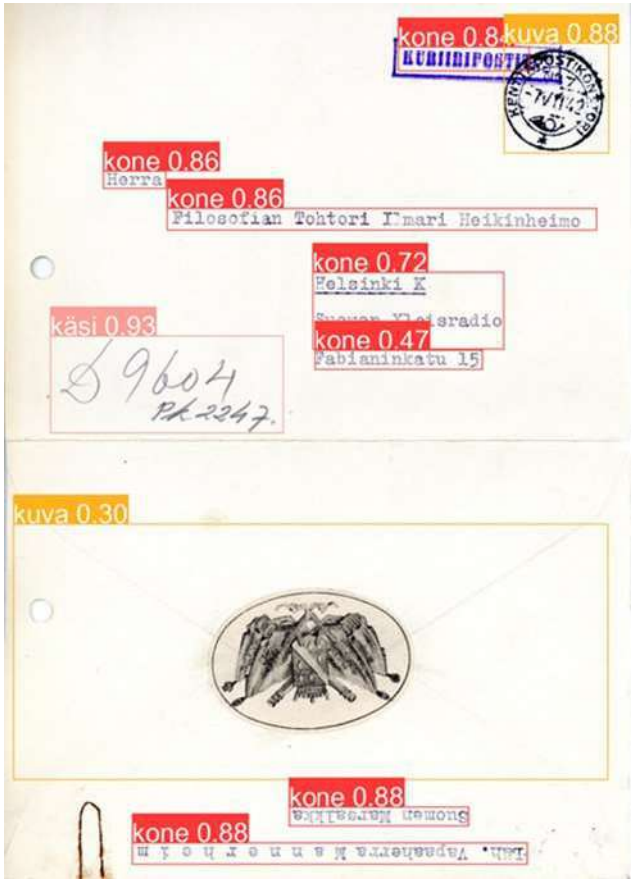
British Council ei järjestä tästä otteesta vaan myöhemmin.

```
Lomake
IPA N:o 401

| Posti- ja lennätinlaitos — Post- och telegrafverket

Mistä *)

Från *) El LONDON LHTSI4 33 4 1015 NORTHRFRN c-

Virkamuistutuksia I

Tjänsteanmärkningar A

Sähkösanoma — Telegram Vastaanotettu — Emottaget

a ELT = MADAMU WUOLIJOKI k ssemetd | [p
E E E VO A 5/9. 6 a
- 8 = Omit A MMUCVV TY T 1! ' ( — e, K:
ta SLATSRADION HLS - i H. 1947 192 k
JN | - E
2:85 Yy * III 7 10 C, p E
4 PILP 4f joht.
jäs ao leün. VO) E

| JN lähetyspaikan nimen jälkeen merkitsevät: 1) sähkösanoman numeroa, 2) sanalukua, 3), 4) ja 5) sähkösanoman sisäänjättö
päivää,
-tuntia ja -minuuttia.

*) Siffrorna efter avsändningsorten betyda: 1) telegrammets nummer, 2) ordantal, 3), 4) och 5) inlämningsdatum, -timme och
-minut.

ROT VA 1 11 N I TOA ra "S1 KA MIIA Pit TIAN FC 1
> UG Tue b f0U MIt An L ANDIIALII U O BIN TA 4 IAIN LO FI
$ 000 + 87 | [5 va TY ÄM TOI MI Tirra 6C'A 12 [6.3 4 % <
RADIO J==)1 [1 2110 J AIN LOTI A LK DA TT 1) 4 1 1)
n TTI TIE m / NA 29=1 H STOP m RR mu) (1
IV CLIN 6 STOCKHO Ki Co = m VOE N TR YU
m - ' ‰» fy (3 J
Ora 3: 1971 n Y ra v AIM MMA PCI ra TY IA S tr 1 AT) 7
Kia PRI OIRM IF LIN LIKI D 4 4 J TNGAG ING HMR KADJIO
, H TAU TAISIN TV A m T A
[4 YT 1419) MI A fftah my M to TII V i PA
U: AAILKSN. TNGAT UMA AN 1 JUSTIEF Y 11C- BALTIC
D (6 C TNT DT TIAN (YY
KUDYU ANT E It RIU JAN J

/ —hrandarmdg
ac / Pe istot Powel

V. e na att a
/ R >? f (punti EA Vary 7 (n ja ha eee dt
) Kin 4 h ( '
```

KANSALLISARKISTO

Muiden sisältöelementtien tunnistus

- Neuroverkkoihin pohjautuvat objektintunnistusmallit mahdollistavat monen tyyppisten elementtien tunnistamisen asiakirjoista
- Valokuvat, piirrokset, kartat, leimat, taulukot, allekirjoitukset…
- Luokittelu: asiakirja sisältää kuvan
- Paikannus: kuva sijaitsee tietyssä kohdassa dokumenttia
- Esimerkiksi asiakirjat sisältävät usein sekä käsin- että konekirjoitettua tekstiä, joiden tulkintaan käytetään omia koneoppimismalleja
    - Kansallisarkistolla testattu käsin- ja konekirjoitettujen tekstikohtien tunnistusta ja erottelemista dokumentista kohdennettua tekstintunnistusta varten



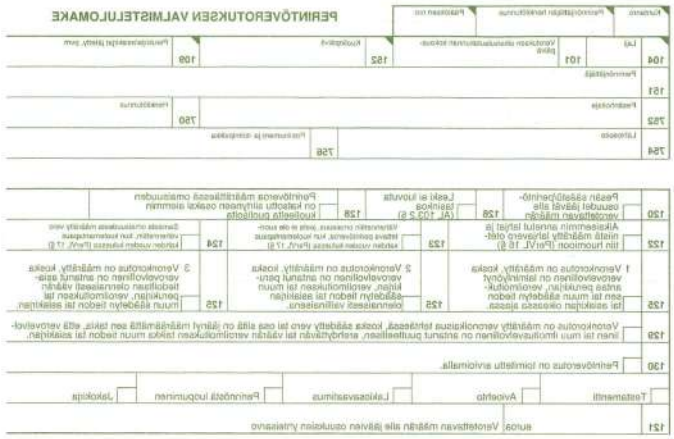
KANSALLISARKISTO

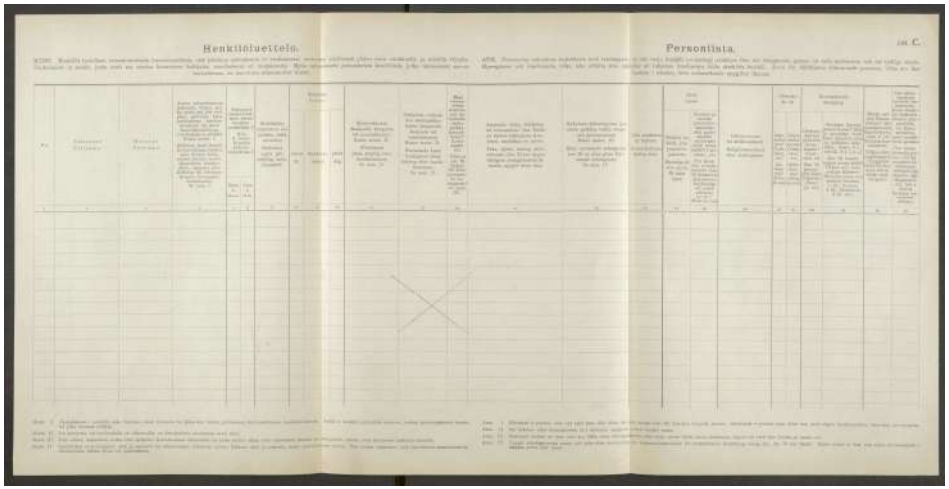
Asiakirjan rakenteiden tunnistus

- Laadukas tekstisisällön tunnistus edellyttää usein rakenteiden tunnistusta: esim. taulukon sisältävän dokumentin syöttäminen suoraan ocr-sovellukselle tuottaa yleensä varsin sekavia tuloksia

- Toisaalta dokumentin rakenteet kuten taulukot ja lomakepohjat halutaan usein muuntaa digitaaliseen muotoon mahdollisimman uskollisesti: tällöin täytyy tunnistaa rakenne ja teksti sekä sijoittaa teksti oikein rakenteeseen

- Rakenteiden tunnistukseen tarjolla erilaisia tekniikoita ja sovelluksia, tässäkään tapauksessa ei ole olemassa yhtä tiettyä dominoivaa menetelmää





KANSALLISARKISTO

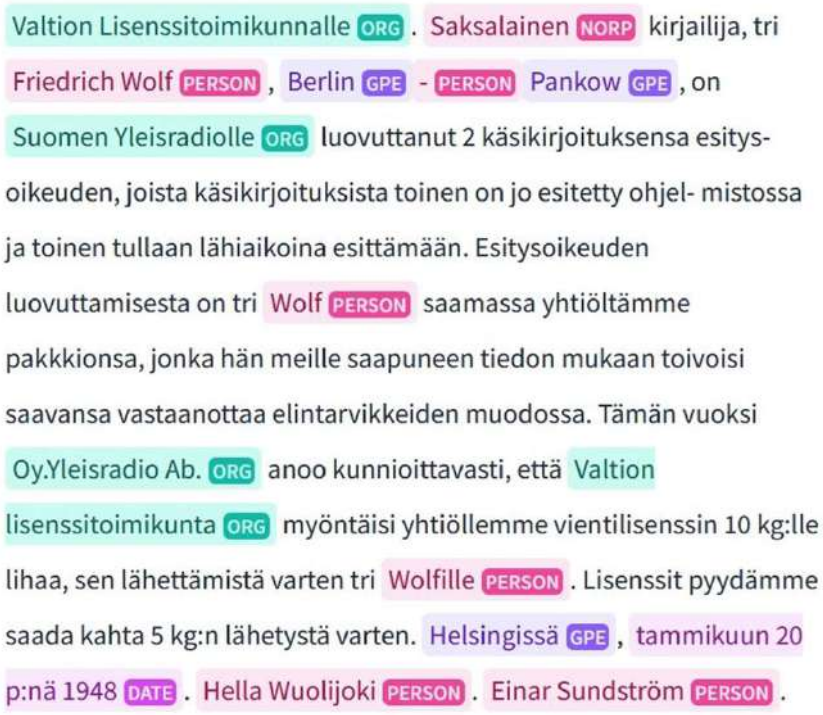
Tunnistetun (teksti)sisällön analysointi

- Aineiston ’raakasisällön’ mahdollisimman laadukas tunnistus on keskeinen vaihe asiakirjojen saattamisessa käyttäjille digitaalisessa muodossa
- Yhä tärkeämmäksi on kuitenkin muodostunut myös tunnistetun sisällön analysointi ja jalostaminen erilaisia käyttötarkoituksia varten
- Käyttäjä voi esimerkiksi haluta etsiä aineistosta tiettyjä henkilöitä tai paikkoja, etsiä tiettyyn paikkaan, tapahtumaan tai aihepiiriin liittyviä asiakirjoja, lukea tiivistelmiä aineistokokonaisuuksien sisällöstä, tai vaikkapa esittää sovellukselle kysymyksiä aineistoon liittyen
- Tässä tekoälyn ja erityisesti suurten kielimallien kehitys tarjoaa monenlaisia mahdollisuuksia

KANSALLISARKISTO

Nimientiteettien tunnistus

- Mistä NERissä on kyse?
  - Tunnistetaan tekstidatasta ennalta määriteltyihin kategorioihin kuuluvia elementtejä
  - Yleisiä entiteettejä esim. henkilön nimet, paikannimet, organisaatiot, aikamääreet..
  - Hyödyllinen työkalu digitoidun aineiston automaattisessa metatiedottamisessa



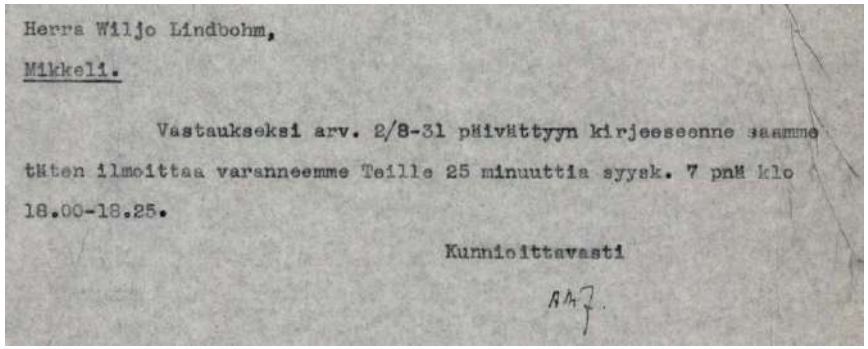
- NER-mallit yleensä koulutettu datalla joka
  a) ei sisällä juurikaan asiakirja-aineistoa ja
  b) ei sisällä tekstintunnistuksen (OCR, HTR) tuottamaa ‘meluista’ aineistoa
- Tästä syystä myös tunnistavat heikosti entiteettejä tällaisesta aineistosta
- Kansallisarkisto toteutti yhdessä Fin-Clariah –tutkimuskonsortion tutkijoiden kanssa erityisesti arkistoaineistoa varten kehitetyn nimientiteettien tunnistusmallin
- Tunnistaa yhteensä 10 luokkaan kuuluvia nimientiteettejä:
  - Henkilönnimet, organisaatiot, paikat, geopoliittiset alueet, tuotteet, tapahtumat, päivämäärät, kansallisuudet ja uskonnolliset tai poliittiset ryhmät, diaarinumerot, y-tunnukset
- Yhdistetyllä arkistoaineistolla sekä uudemmalla (mm. uutistekstejä, blogitekstejä, lakitekstejä) datalla koulutettu malli tunnistaa entiteettejä hyvin eri tyyppisistä aineistoista, hälyisestä OCR:ätystä asiakirjatekstistä tai digisyntyisestä blogitekstistä

```
Pelkkää O
tyhjyyttä       O

Kävin   O
tänään  B-DATE
katsomassa      O
Suomen  B-ORG
Perinteisen     I-ORG
Teatterin       I-ORG
näytelmän       O
Ranta   B-WORK_OF_ART
.       O

Jo      O
teatterin       O
nimi    O
antikliimaksi   O
;       O
ikäänkuin       O
ryhmä   O
tahtoisi        O
antaa   O
katsojalle      O
vakuutuksen     O
siitä   O
```

Esimerkki Turku OntoNotes Entities Corpus-aineistosta

KANSALLISARKISTO

- **Elian Lehto** sijoittui 11:nneksi Kitzbühelin maailmancupin syöksyssä. Yksikään suomalainen mieslaskija ei ole sijoittunut syöksyssä yhtä korkealle maailmancupissa. **Andreas Romar** sijoittui aikoinaan Bormiossa 2012 sijalle 14.


Herra Wiljo Lindbohm,
Mikkeli.

Vastaukseksi arv. 2/8-31 päivättyyn kirjeeseenne saamme täten ilmoittaa varanneemme Teille 25 minuuttia syysk. 7 pnä klo 18.00-18.25.

Kunnioittavasti

AA J.

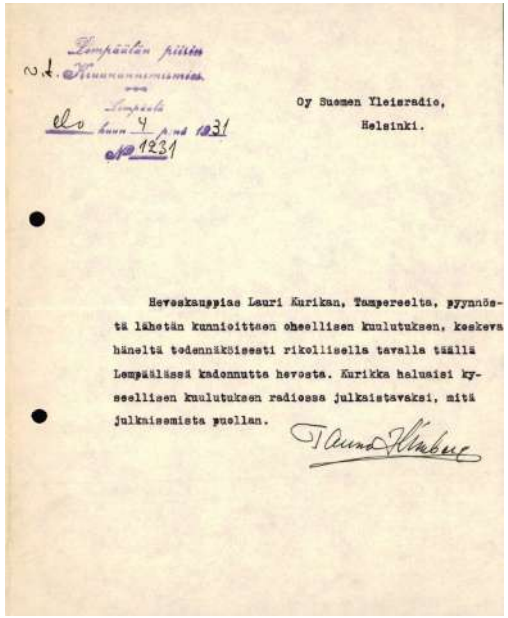
Elian Lehto PERSON sijoittui 11:nneksi Kitzbühelin maailmancupin EVENT syöksyssä. Yksikään suomalainen NORP mieslaskija ei ole sijoittunut syöksyssä yhtä korkealle maailmancupissa. Andreas Romar PERSON sijoittui aikoinaan Bormiossa GPE 2012 DATE sijalle 14.

Mikkeli GPE . , 13 pni elokuuta = 31 DATE ... Herra Wiljo Lindbohm PERSON , N ; N j Vastaukseksi arv. 2/8-31 DATE päivättyyn kirjeeseonne SN S. täten ilmoittaa varanneemme Teille 25 minuuttia syyske. 7 pnä DATE klo a W : 18.00+18.25. | : | i 4 N Kunnioittavasti Eeti W An: dä s R

Automaattinen asiasanoitus

- Nimientiteettien tunnistusmalli etsii asiakirjoista vain niitä ennalta määriteltyjä entiteettejä joilla malli on koulutettu
  - Ei ole yleensä paras vaihtoehto aineiston sisällön kuvailuun
- Automaattinen asiasanoitus tuottaa syötteenä saadun tekstin pohjalta sitä kuvailevia asiasanoja
- Tunnettu suomenkielinen sovellus Kansalliskirjaston toimesta kehitetty **Annif** (https://annif.org/)
  - Hyödyntää koneoppivia neuroverkkomalleja ja asiasanalistoja (esim. YSO - Yleinen suomalainen ontologia)
  - Annifia käyttää mm. automaattinen sisällönkuvailun palvelu Finto AI (https://ai.finto.fi)

Automaattinen asiasanoitus



Oy Suomen Yleisradio,
Helsinki.

Hevoskauppias Lauri Kurikan, Tampereelta, pyynnöstä lähetän kunnioittaen oheellisen kuulutuksen, koskeva häneltä todennäköisesti rikollisella tavalla täällä Lempäälässä kadonnutta hevosta. Kurikka haluaisi ky- seellisen kuulutuksen radiossa julkaistavaksi, mitä julkaisemista puollan.

PA
(W 4 6 €.) 6 04 64 91 4678 906071 6Ä 199664,
Oy Suomen Yleisradio,
«i tmÄ Kala
o kuun A A:nd 1931 Helsinki.

Hevoskauppias Lauri Kurikan, Tampereelta, pyynnöstä lähetän kumioittaen oheellisen kuulutuksen, koskeva häneltä todennäköisesti rikollisella tavalla täällä Lempäälässä kadonnutta hevosta. Kurikka haluaisi ky- seellisen kuulutuksen radiossa julkaistavaksi, mitä julkaisemista puollan. ET A
/ v

Get suggestions → annif

SUGGESTED SUBJECTS

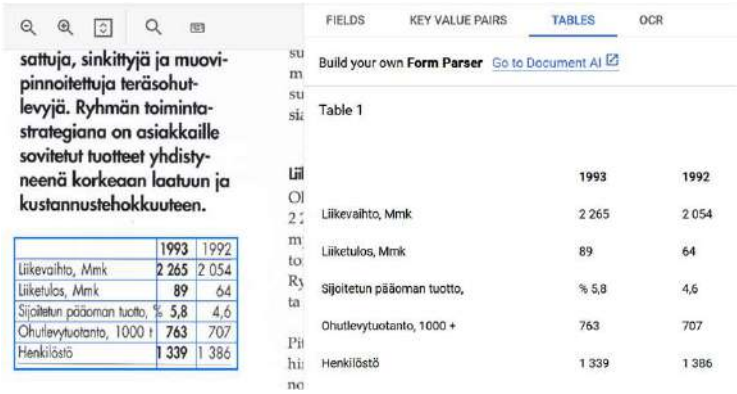
- historia
- Kurikka
- Lempäälä
- suomen kieli
- muistelmat
- yleisradiotoiminta
- radio-ohjelmat
- paikallishistoriat (historiikit)
- hevonen
- paikannimet

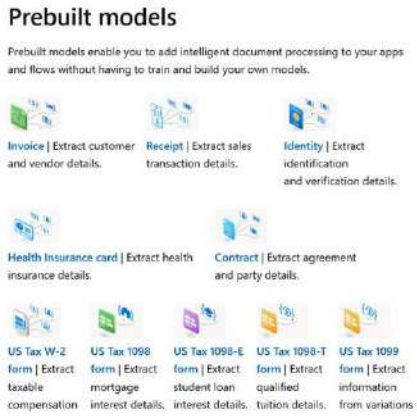
Digitoitu asiakirja, tekstintunnistuksen (Tesseract OCR) siitä tunnistama tekstisisältö sekä automaattisen asiasanoituksen (Annif) tekstin pohjalta tuottamat asiasanat.

KANSALLISARKISTO

3. Käytettävyys: miten tekoälysovelluksiin pääsee käsiksi?

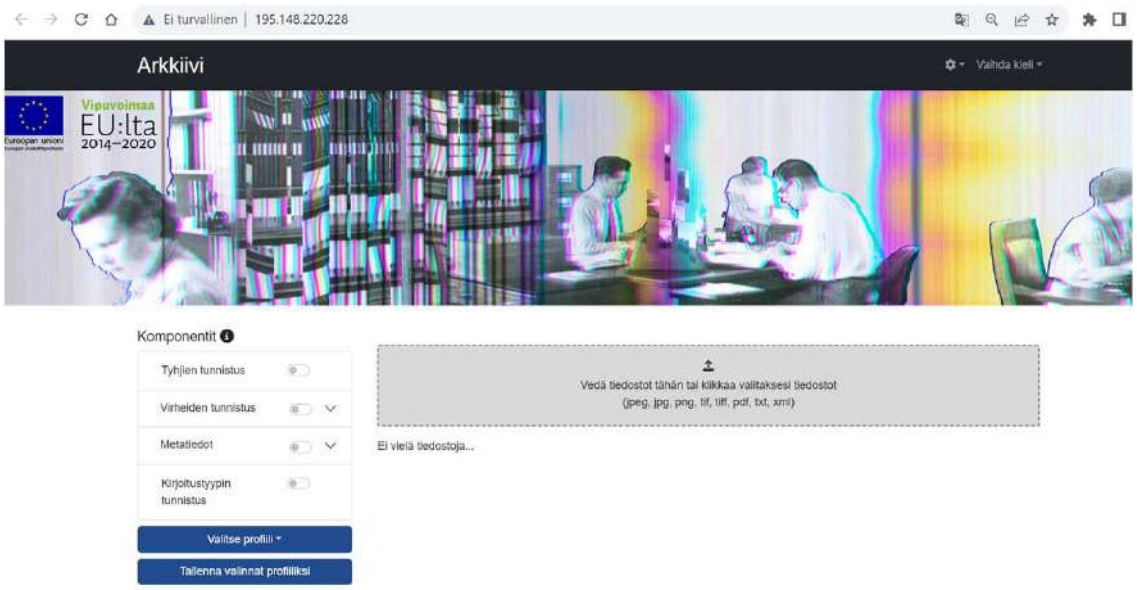
- Suurilla pilvipalvelun tarjoajilla on omat alustansa joissa saatavilla erilaisia (pääosin maksullisia) sovelluksia asiakirjojen automaattiseen prosessointiin. Esimerkiksi
  - Amazon Textract: kone- ja käsinkirjoitetun tekstin ja dokumentin rakenteiden tunnistus
  - Amazon Comprehend: kielen tunnistus, (tiettyjen) nimientiteettien tunnistus, dokumenttien luokittelu (positiivinen / negatiivinen sisältö)…
  - Microsoft Azure AI Document Intelligence: dokumentin tekstisisällön ja rakenteiden tunnistus…
  - Microsoft Azure AI Language: (tiettyjen) nimientiteettien tunnistus, dokumenttien luokittelu (positiivinen / negatiivinen sisältö), tiivistelmien teko…
  - Google Document AI: kone- ja käsinkirjoitetun tekstin ja dokumentin rakenteiden tunnistus…





KANSALLISARKISTO

- Kaupallisten pilvipalveluiden haasteena mm. maksullisuus, englannin dominanssi (NLP-sovellukset), pilviprosessoinnin tietosuojahaasteet..
    - Toisaalta isojen toimijoiden resurssit mahdollistavat myös laadukkaiden sovellusten tuottamisen (esim. OCR)
- Myös ilmaisia, suomenkieliselle aineistolle suunnattuja sovelluksia saatavilla lisääntyvässä määrin
- **Arkkiivi** (https://arkkiivi.fi/):
    - Alusta sisältää komponentteja jotka tunnistavat digitoiduista aineistoista skannausvirheitä ja sisältöjä
    - Demo/kokeilualusta, joka ei sovellu tuotantokäyttöön: tätä varten sovellusten koodit vapaasti saatavilla GitHubissa



4. Tekoäly Kansallisarkistolla

- 2016-2019 EU-rahoitteinen READ-hanke (Recognition and Enrichment of Archival Documents): Transkribus
- Sisällönanalysointiprojekti 2018-2019 – OCR:aan ja digitoinnin tekoälytyökalujen selvityshanke (DALAI:n esihanke)
- 2020 – Tuomiokirjahaku (3,1 miljoonaa sivua sisältötunnistettuna)
- 1.9.2021–31.8.2023 DALAI (Digitaalisten aineistojen laadun ja käytettävyyden parantaminen tekoälyavusteisesti)
- 2021 – Pohjoismainen yhteistyö tiivistyy (hackathonit, kuukausittaiset tekoälykehittäjien verkkotapaamiset)
- 2022 -> Oman tekoälykyvykkyyden nosto. Sisällöntunnistuksen ottaminen omiin käsiin ja muiden menetelmien tutkiminen.
- 2023 Uusi Tutkimus ja innovaatiot –toiminto vastaamaan tekoälyn hyödyntämisestä
    - 3 koneoppimisen asiantuntijaa, 5 arkistoalan ja tutkimuksen asiantuntijaa

KANSALLISARKISTO

Käsinkirjoitetun tekstin tunnistus

- **HTR** (Handwritten Text Recognition) ollut pitkään haastava osa-alue, käsialojen ja käsinkirjoitetun aineiston monipuolisuuden takia vaikea kehittää yleisiä tekstintunnistusmalleja
- Ei juurikaan ilmaisia sovelluksia, vaaditaan usein omien mallien kouluttamista tiettyjä aineistokokonaisuuksia varten
- Arkistojen, kirjastojen ja yliopistojen EU-projektin yhteydessä perustama **Transkribus** (https://readcoop.eu/transkribus) on tunnettu ja monien kulttuuriperintöalan toimijoiden käyttämä alusta dokumenttien annotointiin, käsialamallien kouluttamiseen ja hyödyntämiseen
    - Kansallisarkisto ollut myös mukana perustamisesta asti, mutta korkeiden kustannusten takia siirrytty omien HTR-mallien kehittämiseen

Käsinkirjoitetun tekstin tunnistus

- Esimerkki tekstintunnistuksen eri työvaiheista

Tekstialueen tunnistus

- Alkuperäinen kuva

Tekstirivin tunnistus

- Alkuperäinen kuva

Käsinkirjoitetun tekstin tunnistus

- Rivikuvat

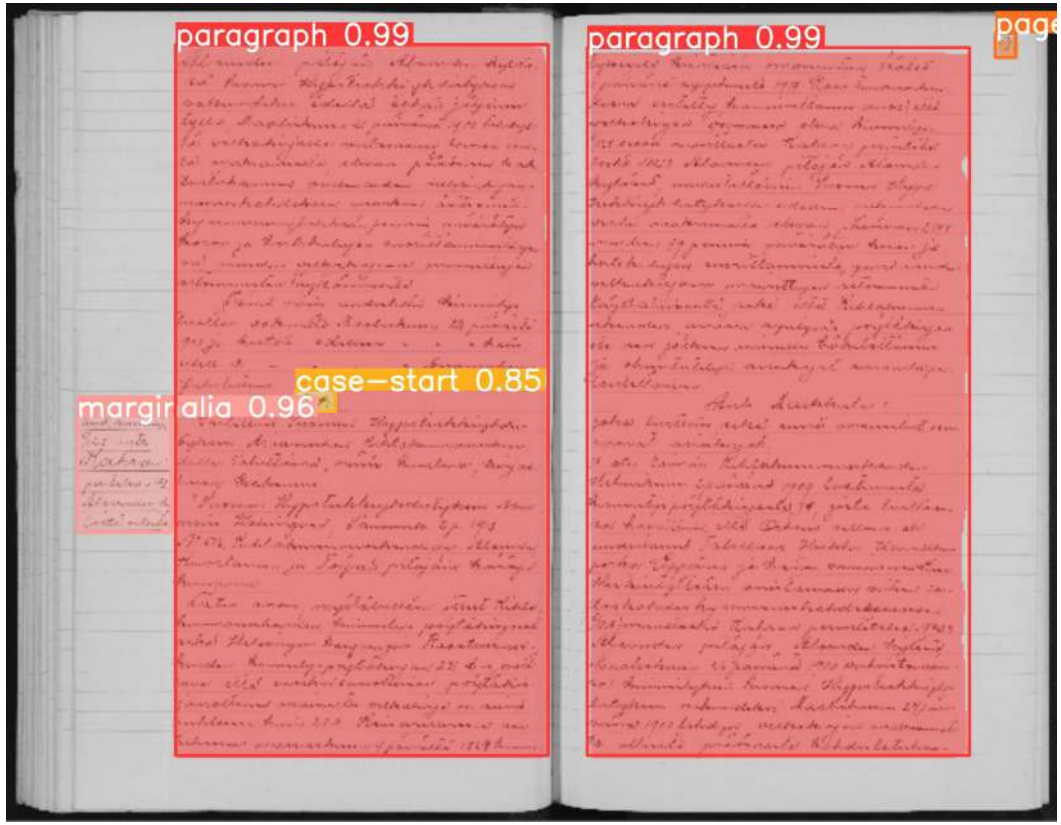
ALTO XML-muunnos

- Tunnistettu teksti
- Tunnistetut tekstialueet
- Tunnistetut tekstirivit

KANSALLISARKISTO

Tekstialueiden tunnistus

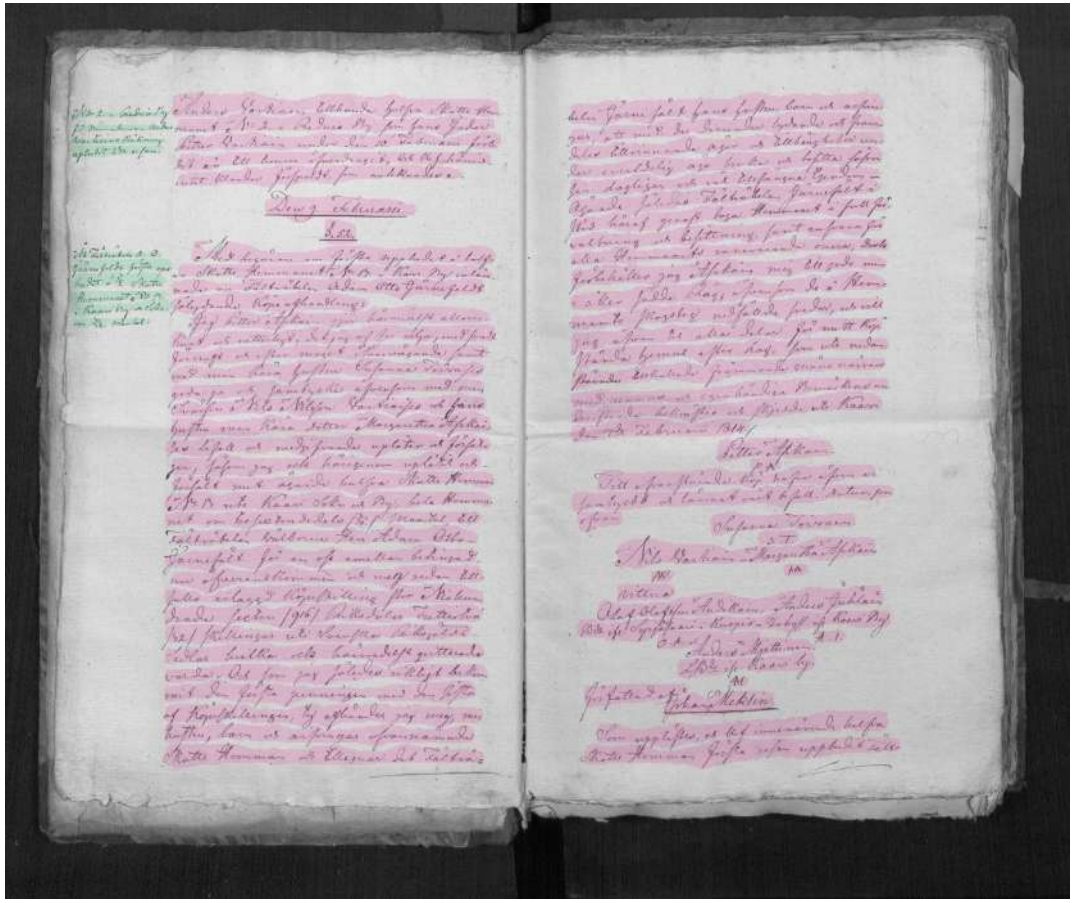
- Tunnistetaan sivulta tai aukeamalta eri tyyppisiä kokonaisuuksia
- Tuomiokirjaesimerkissa malli tunnistaa neljän tyyppisiä alueita:
  - Tekstikappale
  - Marginaalissa oleva teksti
  - Sivunumero
  - Uuden oikeusjutun alku



KANSALLISARKISTO

Tekstirivien tunnistus

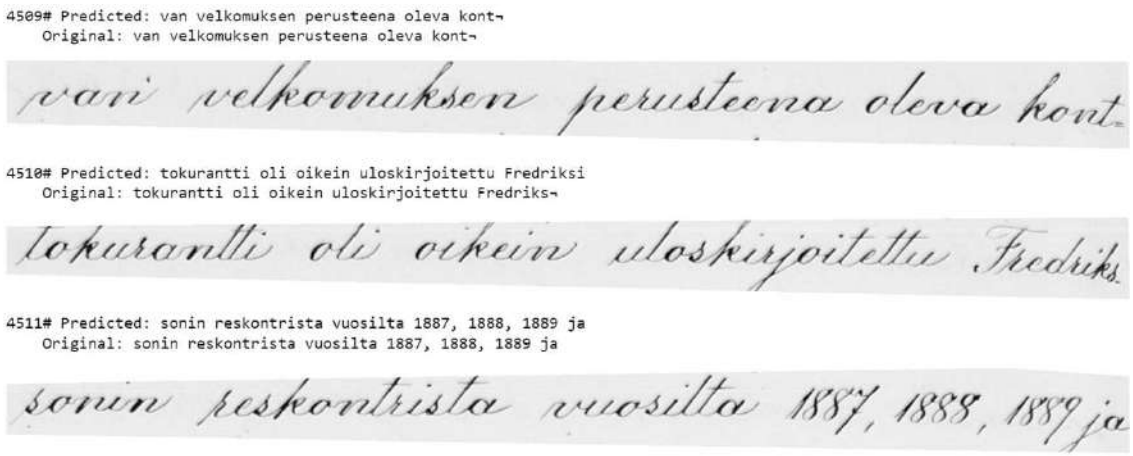
- Tunnistaa dokumentista tekstirivit
- Rivit yhdistetään koordinaattien perusteella aiemmin tunnistettuihin tekstialueisiin
- Hyödynnetään objektintunnistukseen erikoistunutta neuroverkkomallia



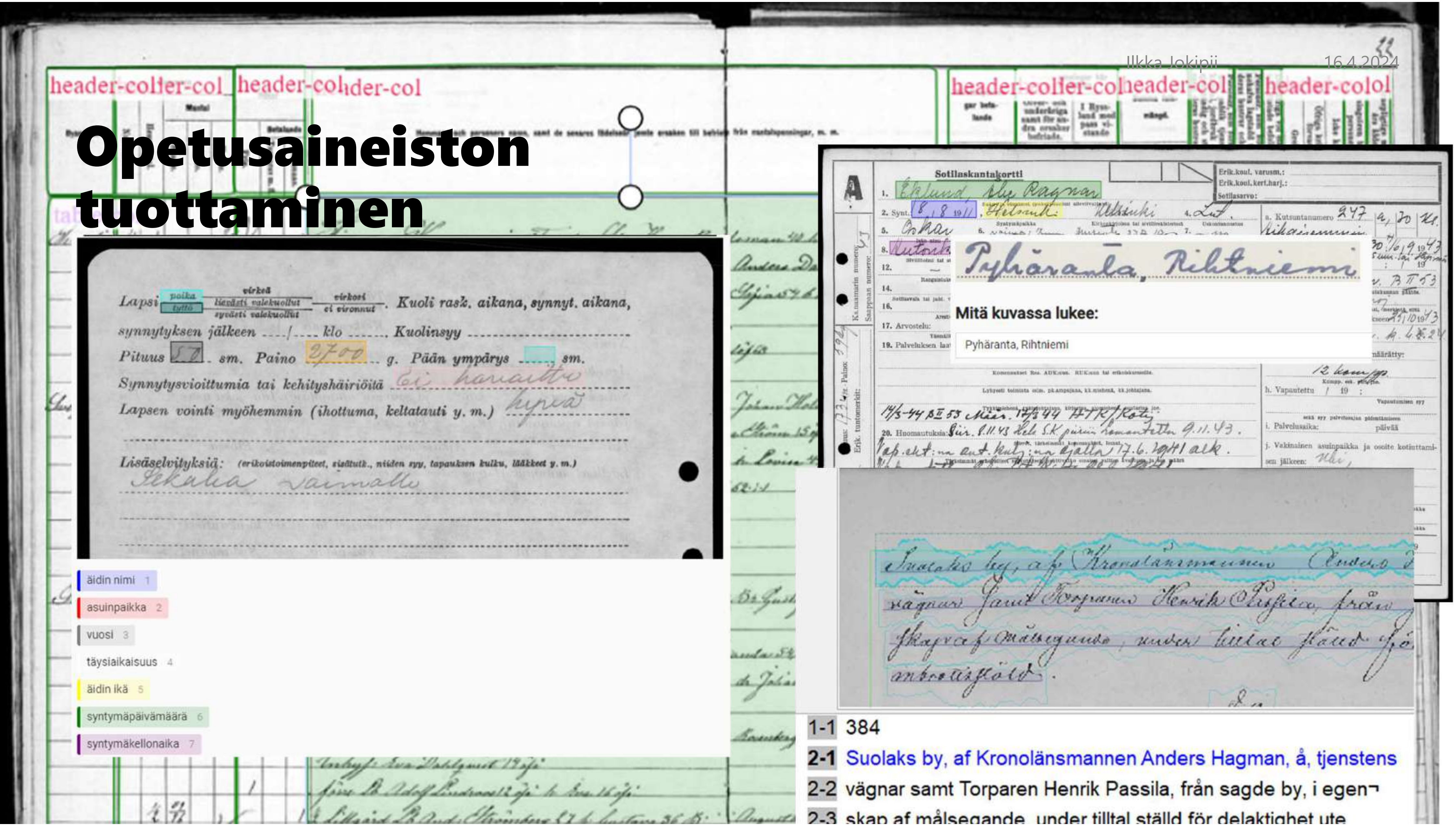
KANSALLISARKISTO

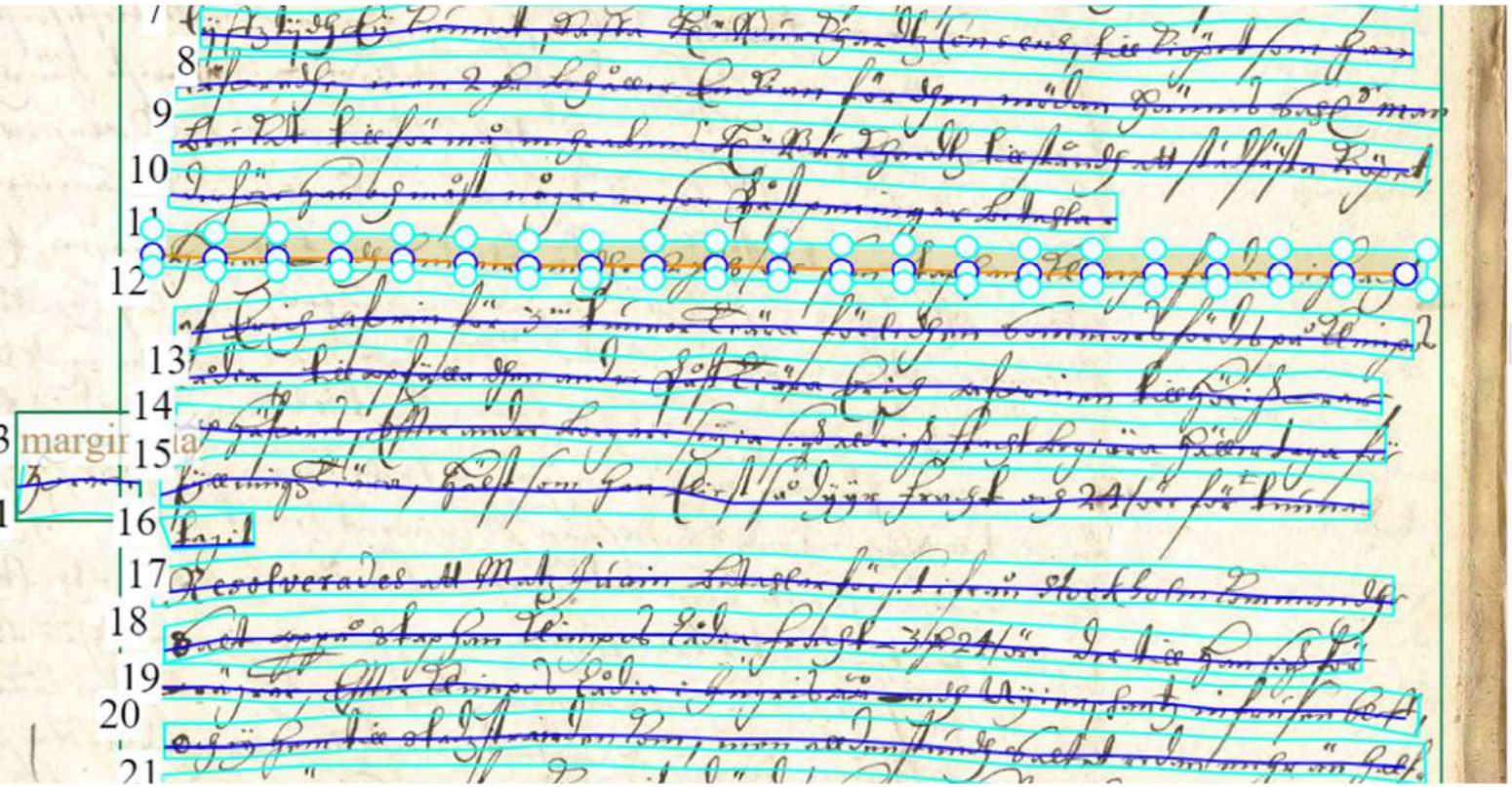
Tekstisisällön tunnistus

- Tekstintunnistusmalli saa syötteenä rivikuvia ja tunnistaa niistä tekstisisältöä
- Neuroverkkomallin koulutusdatana käytetty yli 300 000 rivikuvaa, joihin on käsin annotoitu tekstisisältö



Opetusaineiston tuottaminen





7 Enckia betahla tillbakas åth Pafwilla 3 D:r, effter Sahl:gh Torfwinen i sin

8 lijffztijdh Eij kunnat skaffa H:r Burkhardtz Consens, till kiöpet som han

9 låfwadhe, men 2 D:r behåller Enckian för dhen mödan hännes Sahl:ge Man

10 brukat till förmå mehr bem:e H:r Burkhardtz tillståndh att städfästa kiöpet,

11 derföre Han och måst några reesor Påstpeningar betahla.

12 Huadh wedh kommer om dhe 2 D:r 8 / öre som Staphan Klimpo fordrar i fracht

13 af Erich Afwin för 3:ne tunnor Tiära förledhen Sommars fördes på Klimpos

14 Lådia, till upfylla dhen andre Påst Tiära Erich Afwinen till hörigh war

15 ophäfwes, effter andre borgare seijia sigh aldrigh fracht begiära häller taga för

16 fyllningh Tiära, hälst som han Eliest så dyyr fracht och utfööre för tunna

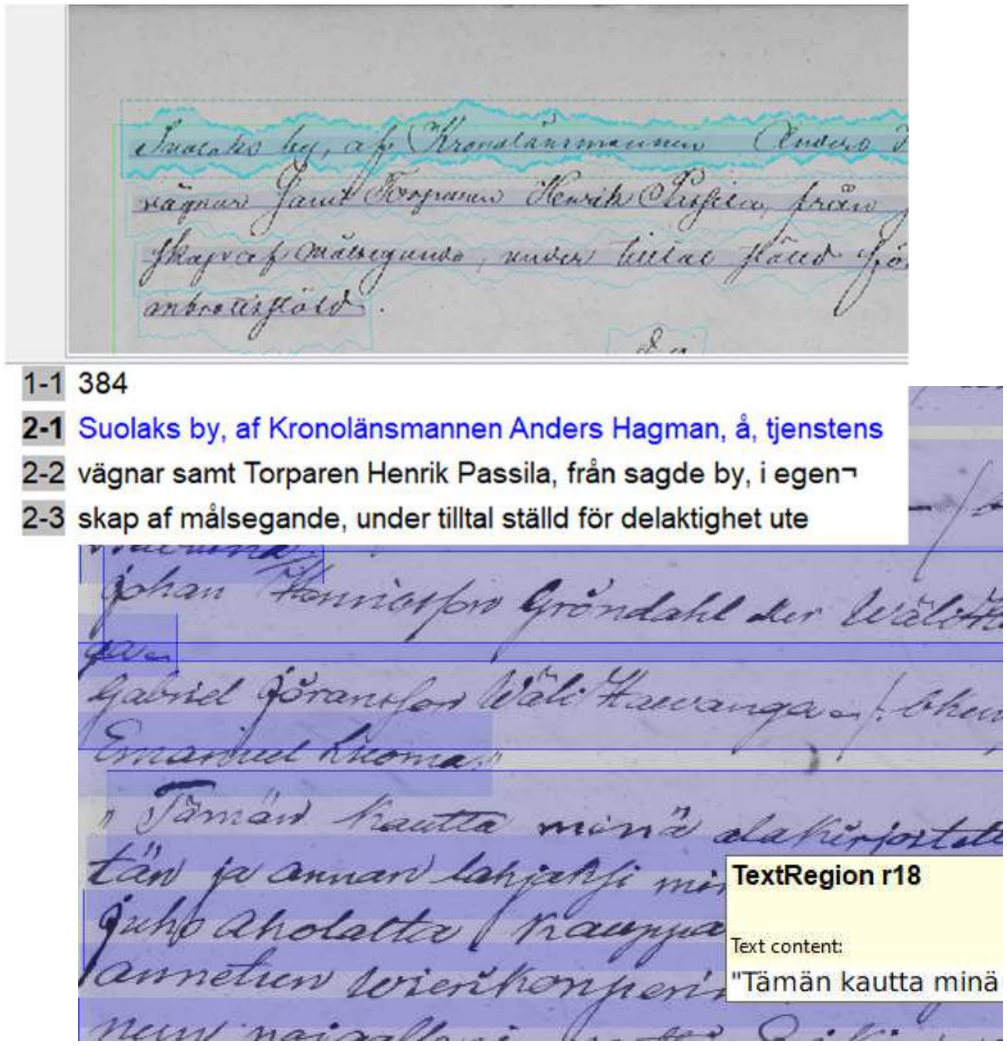
17 tagit.

18 Resolverades att Matz Juoin betahlar för sit ifrån Stockholm kommandhe

KANSALLISARKISTO

Kansallisarkiston opetusaineistokokonaisuuksia

- READ-hanke & 1800-luvun tuomiokirjat: n. 1,3 miljoonaa sanaa
- Joukkoistushankkeet: n. 2 miljoonaa sanaa
  - Sotapäiväkirjoja
  - 1800-luvun perukirjoja
  - Henkikirjoja
- Muut arkistot ja tutkijat: n. 3 miljoonaa sanaa
  - Riksarkivetilta 1600- ja 1800-lukujen aineistoa
  - Ville Pekka Kääriäinen (HY) merkittävä määrä 1600-lukua
  - Tukholman kaupunginarkistolta 1700-luvun aineistoja

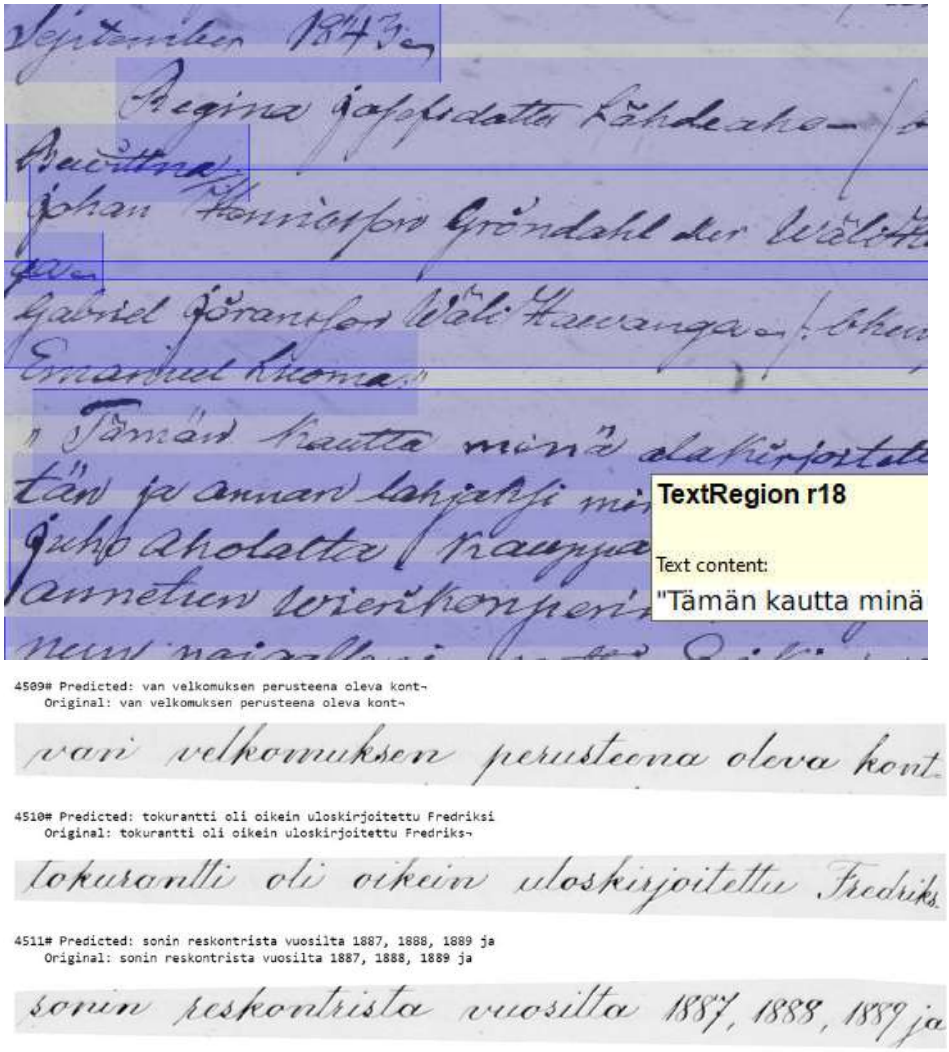


KANSALLISARKISTO

Tuomiokirjat-malli

- Tuomiokirjojen käsialamalli perustuu kahteen tekoälymalliin
- Segmentointi: YOLOv8
  - Parannettu segmentointia, mm. juttujen alkujen tunnistus
- Tekstintunnistus: TrOCR
  - Microsoftin kehittämä käsiala-aineistolla esikoulutettu malli
  - Opetusaineistoa yli miljoona sanaa

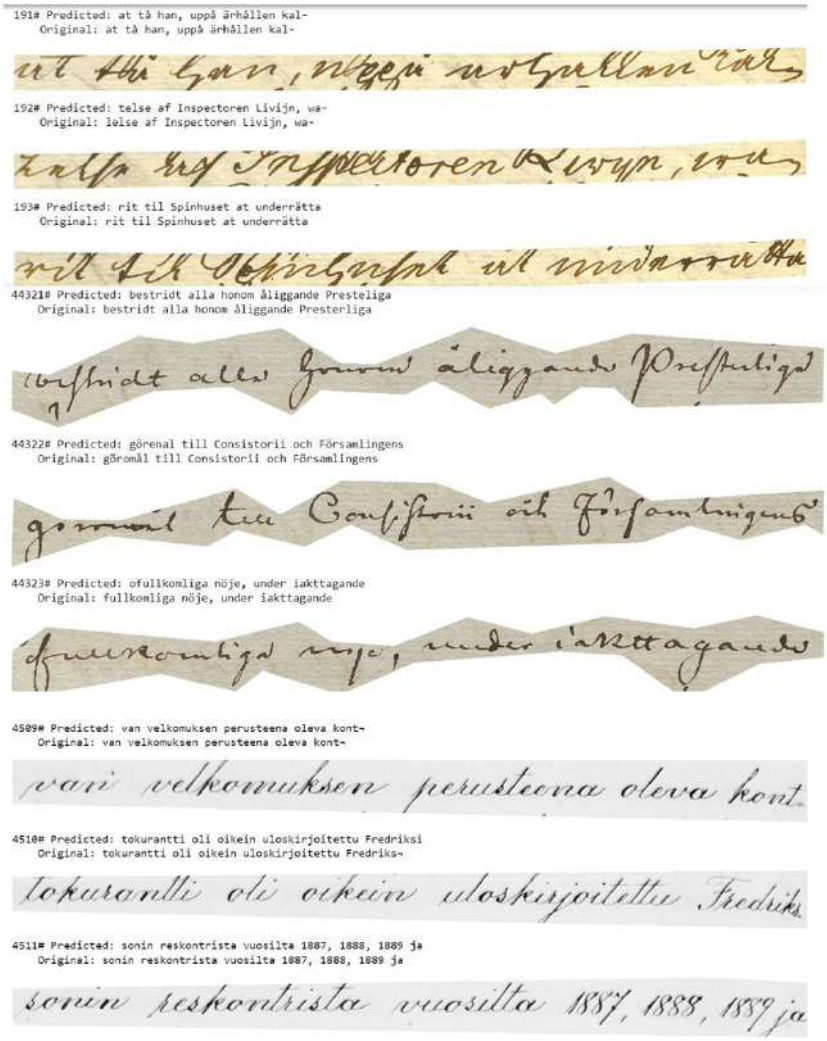
Merkkivirhe 2,4 %



KANSALLISARKISTO

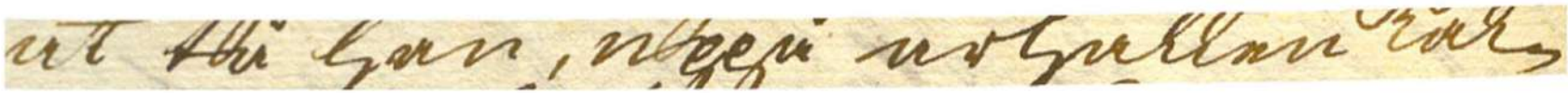
Suomalainen supermalli v. 0.1

- Viimeisin kokeilu käsialantunnistukseen liittyen
- Käsialamalli, joka lukee tekstiä 1600-luvulta 1920-luvulle
    - Koulutettu yhteensä 737 000 rivillä
    - Laskenta kesti reilun kaksi viikkoa
    - Virheprosentti 3,66
- Opetusaineistoa kerättynä huomattava määrä lisää
    - Supermallin seuraava versio vuoden 2024 aikana
        - Huomattava määrä lisäaineistoa -> tarkkuuden parantuminen
        - Jatkokoulutus 1500-luvulle?
- Supermalli toimii myös pohjamallina
    - Koulutettu useita aineistokohtaisia käsialamalleja supermallin päälle -> ei tarvita kovin paljoa uutta aineistoa



KANSALLISARKISTO

191# Predicted: at tå han, uppå ärhållen kal-
Original: at tå han, uppå ärhållen kal-

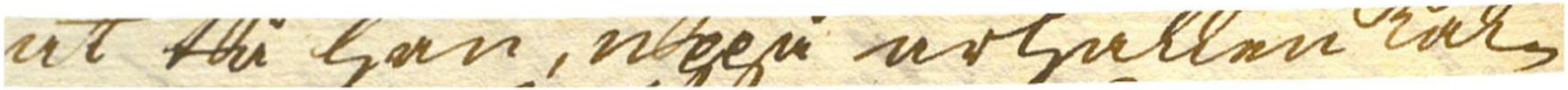


192# Predicted: telse af Inspectoren Livijn, wa-
Original: lelse af Inspectoren Livijn, wa-

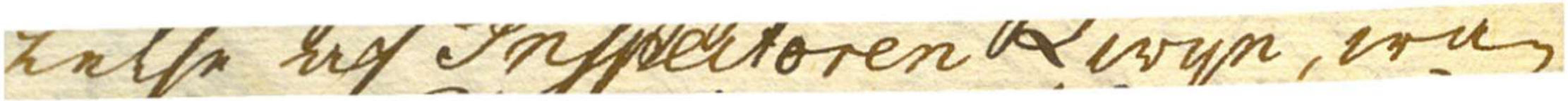
193# Predicted: rit til Spinhuset at underrätta
Original: rit til Spinhuset at underrätta

KANSALLISARKISTO

191# Predicted: at tå han, uppå ärhållen kal-
Original: at tå han, uppå ärhållen kal-

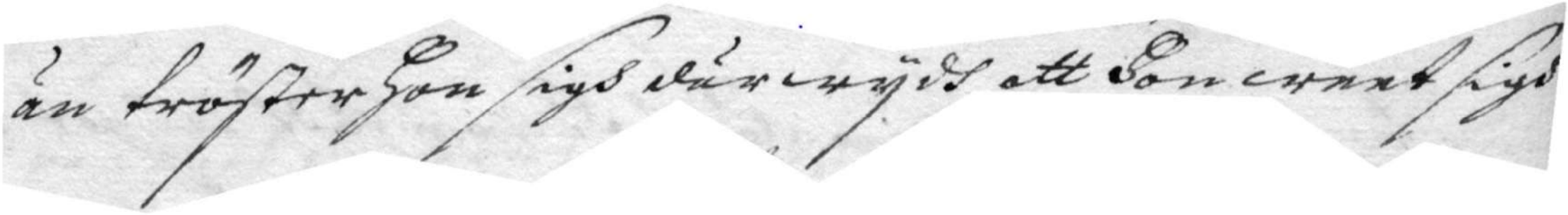


192# Predicted: telse af Inspectoren Livijn, wa-
Original: lelse af Inspectoren Livijn, wa-



193# Predicted: rit til Spinhuset at underrätta
Original: rit til Spinhuset at underrätta

83# Predicted: än tröster hon sigh där wydh att hon weet sigh
Original: än tröster Hon sigh därwydh att Hon weet sigh



84# Predicted: wara oskylldigh, och tackar Gudh därföre
Original: wara oskyldigh, och tackar Gudh därföre

KANSALLISARKISTO

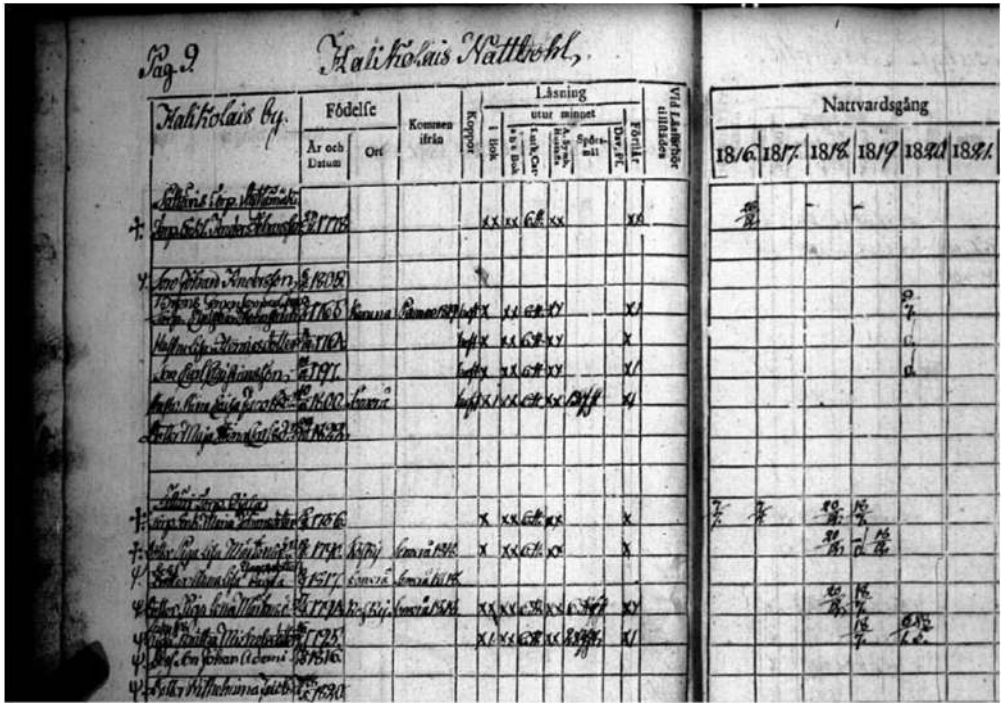
1500-luku!

GT: Riddare, att han medh edher trogne undersåtherr, sá
Pred: Riddare, att han medh edher wogne undersåther, så
CER :0.059

GT: tree eller fyre präster, effter som the haffue gelden och
Pred: rer eller fyra präster, effter som dhe haffue gatan och
CER :0.14

Rakenteelliset aineistot

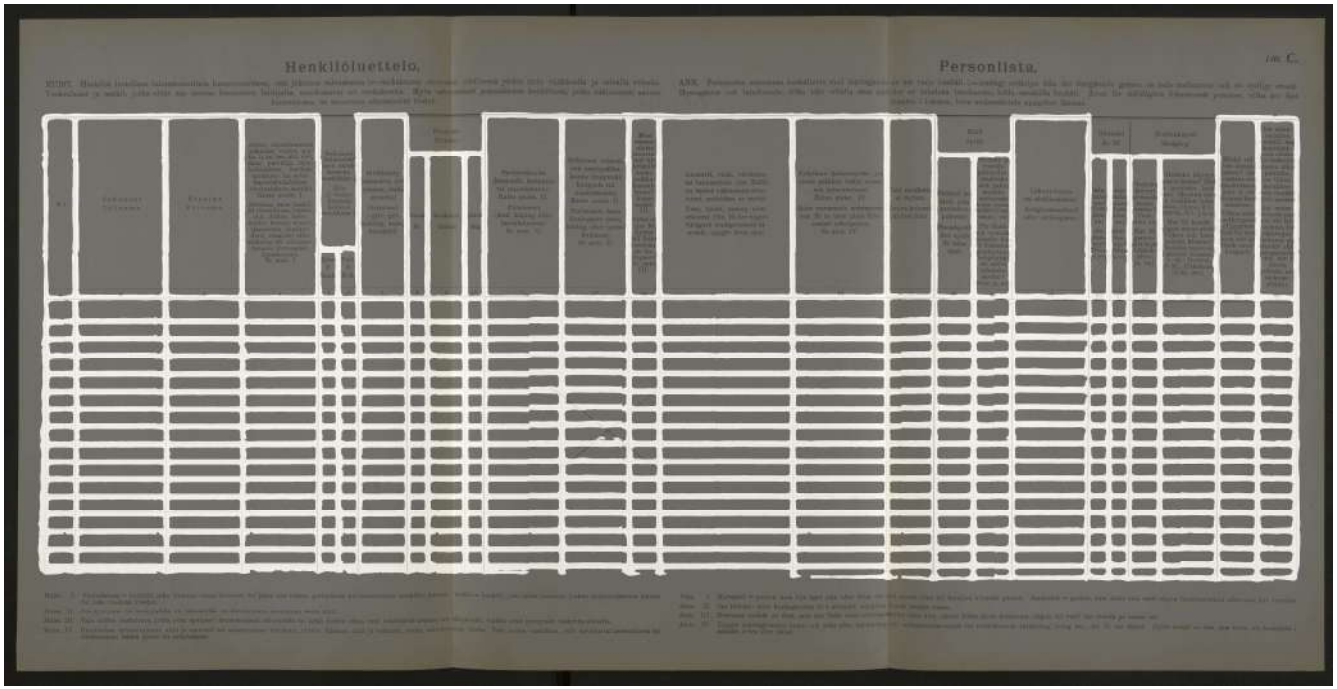
- Kansallisarkisto täynnä erilaista vanhaa tilasto- ja rekisteriaineistoa
  - Henkikirjat ja kirkonkirjat
  - Väestönlaskentatietoja
  - Kuolinsyytilastot
  - Laivojen lokikirjoja
  - Kiinteistörekisterit
- Aineistojen koneellinen tunnistus tarjoaa huikeat mahdollisuudet tuottaa aineistoja laajasti eri tieteenaloille
- Taulukkorakenteiden tunnistus ollut pitkään ongelma tekoälylle, mutta keinoja alkaa nyt löytyä



KANSALLISARKISTO

Taulukkomaskin tunnistus

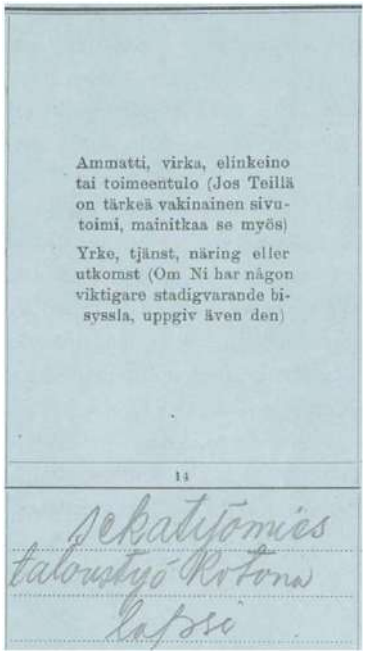
- Annotoiduilla taulukkokuvilla koulutettu neuroverkkomalli tunnistaa kuvasta taulukkorakenteen
    - Vaatii 200-300 annotoitua taulukkoa toimiakseen
- Lopputuloksena taulukkosolujen koordinaatit



KANSALLISARKISTO

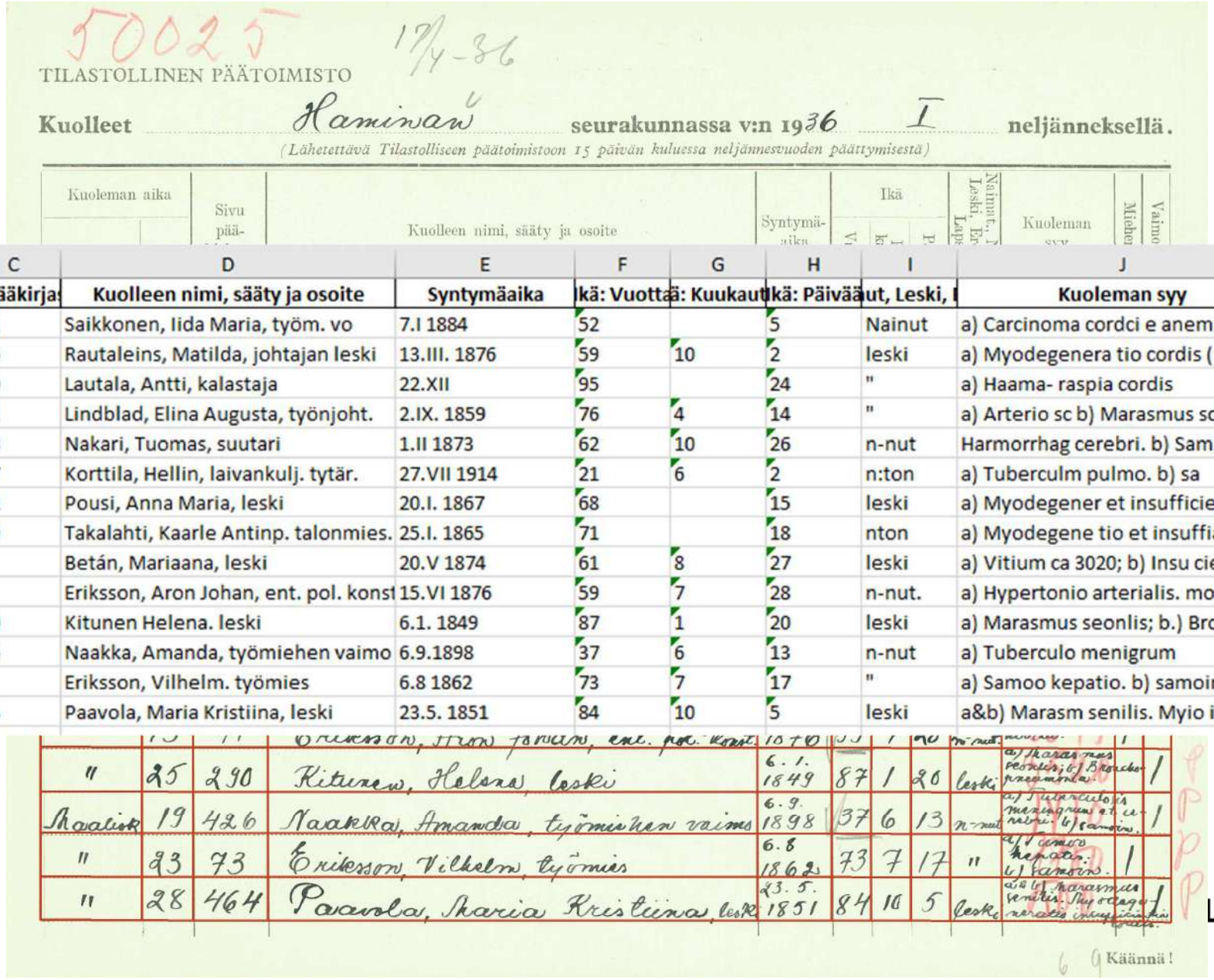
Tekstin tunnistus taulukkosoluista

- Taulukkokoordinaattien perusteella solut voidaan leikata omiksi kuvikseen
- Solukuvat toimivat syötteenä tekstintunnistusmallille
- Tuloksena ennustettu tekstisisältö, joka voidaan viedä esim. excel-taulukkoon



Ammatti, virka, elinkeino tai toimeentulo?
sekatyömies
taloustyö kotona
1 lapsi

KANSALLISARKISTO



50025 17/4-37

TILASTOLLINEN PÄÄTOIMISTO

Kuolleet Kaminan seurakunnassa v:n 1936 I neljänneksellä.

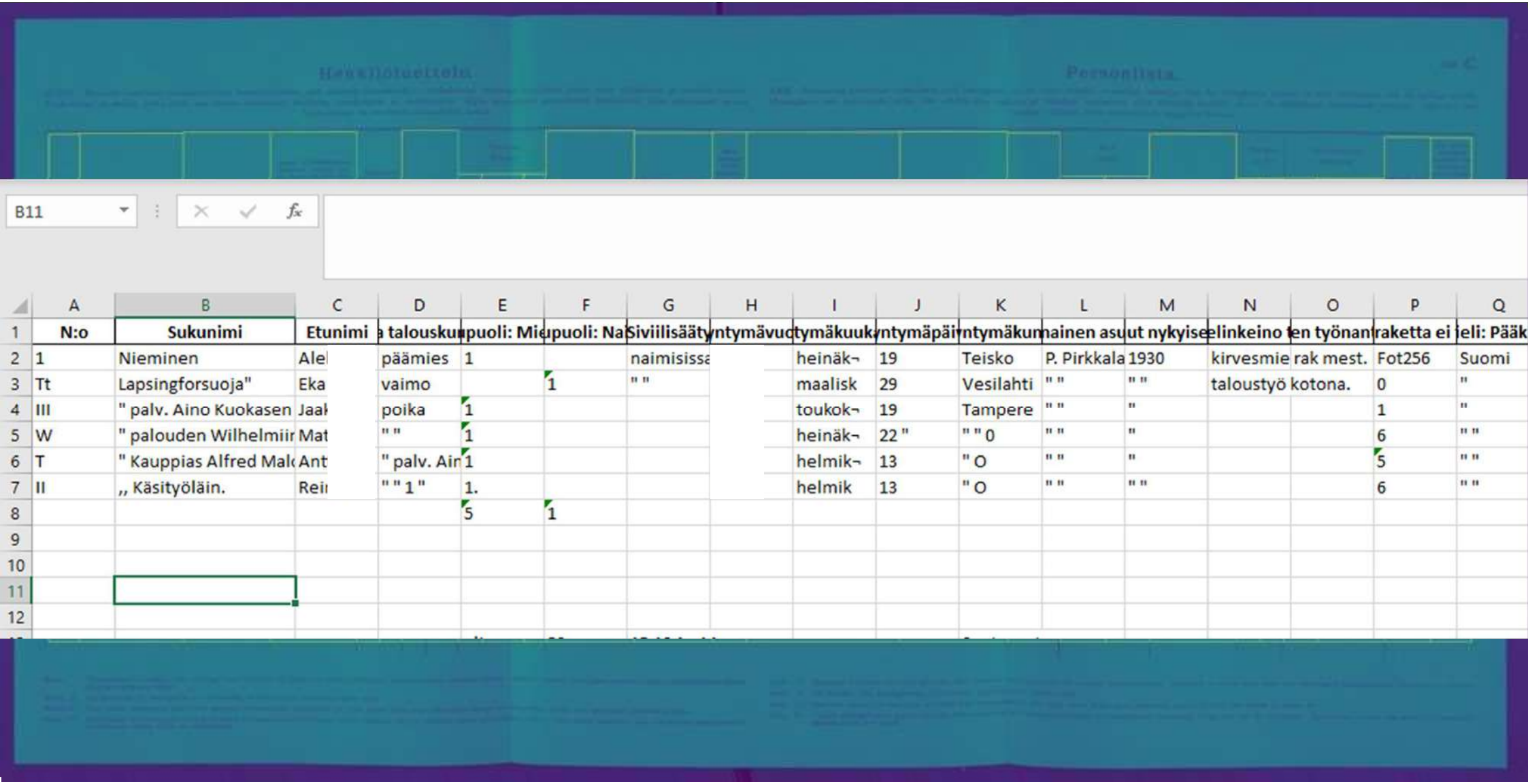
(Lähetettävä Tilastolliseen päätoimistoon 15 päivän kuluessa neljännesvuoden päättymisestä)

Kuoleman aika	Sivu pää-	Kuolleen nimi, sääty ja osoite	Syntymä- aika	Ikä			Nainut, Leski, Er	Kuoleman syy	Mieh	Vaimo
				V.	k.	P.				
11	25	290	Kituinen, Helena, leski	6.1. 1849	87	1	20	leski	a) Marasmus seonlis; b.) Bronchopneumonia	1
Maalisk	19	426	Naakka, Amanda, työmiehen vaimo	6.9. 1898	37	6	13	n-nut	a) Tuberculo is meningrum et. u- ratii. b) samoin.	1
11	23	73	Eriksson, Vilhelm, työmies	6.8 1862	73	7	17	11	a) Samoo hepatio. b) samoin.	1
11	28	464	Paavola, Maria Kristiina, leski	23. 5. 1851	84	10	5	leski	a&b) Marasmus senilis. Myio insuf	1

Käännä!

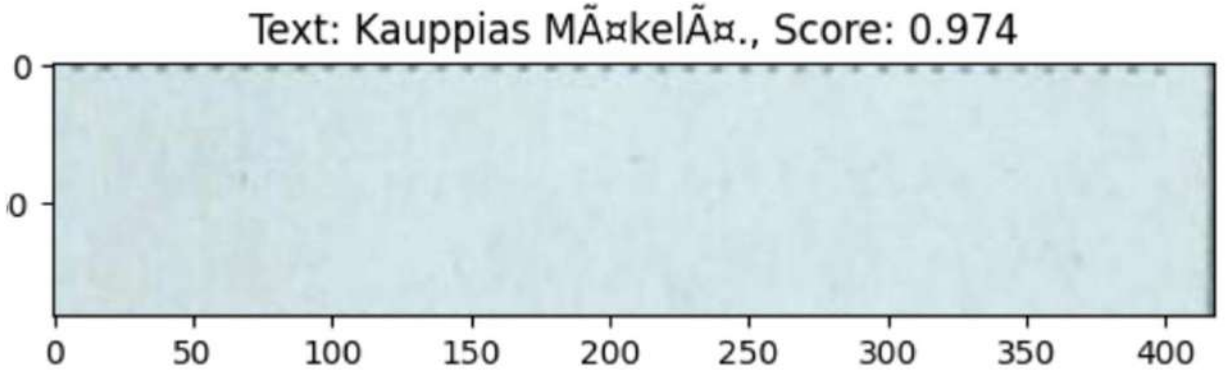
	A	B	C	D	E	F	G	H	I	J	K	L
1	Kuukausi	Päivä	u pääkirja	Kuolleen nimi, sääty ja osoite	Syntymäaika	kä: Vuotta	ä: Kuukaut	kä: Päivää	ut, Leski, I	Kuoleman syy	Miehenp.	Vaimonp.
2	Tammik.	12	591	Saikkonen, Iida Maria, työm. vo	7.I 1884	52		5	Nainut	a) Carcinoma cordci e anemia secun	dis	1
3	"	15	546	Rautaleins, Matilda, johtajan leski	13.III. 1876	59	10	2	leski	a) Myodegenera tio cordis (3030 sar	81	lis
4	"	16	920	Lautala, Antti, kalastaja	22.XII	95		24	"	a) Haama- raspia cordis	1	
5	"	20	371	Lindblad, Elina Augusta, työnjoht.	2.IX. 1859	76	4	14	"	a) Arterio sc b) Marasmus sc	eron	1
6	"	27	428	Nakari, Tuomas, suutari	1.II 1873	62	10	26	n-nut	Harmorrhag cerebri. b) Samoin.	19/	
7	"	29	297	Korttila, Hellin, laivankulj. tytär.	27.VII 1914	21	6	2	n:ton	a) Tuberculm pulmo. b) sa	s. mini	1
8	Helmik.	5	512	Pousi, Anna Maria, leski	20.I. 1867	68		15	leski	a) Myodegener et insufficien cordis	atio tio si	1
9	"	13	689	Takalahti, Kaarle Antinp. talonmies.	25.I. 1865	71		18	nton	a) Myodegene tio et insuffia cordia	sa 7/5 ole	in.
10	"	17	44	Betán, Mariaana, leski	20.V 1874	61	8	27	leski	a) Vitium ca 3020; b) Insu cientia co	Pelin	1
11	"	13	71	Eriksson, Aron Johan, ent. pol. konst	15.VI 1876	59	7	28	n-nut.	a) Hypertonio arterialis. mon	1	
12	"	25	290	Kitunen Helena. leski	6.1. 1849	87	1	20	leski	a) Marasmus seonlis; b.) Bro pneum	velko-	1
13	Maalisk	19	426	Naakka, Amanda, työmiehen vaimo	6.9.1898	37	6	13	n-nut	a) Tuberculo menigrum	sia. bää in	1
14	"	23	73	Eriksson, Vilhelm. työmies	6.8 1862	73	7	17	"	a) Samoo kepatio. b) samoin.	1	
15	"	28	464	Paavola, Maria Kristiina, leski	23.5. 1851	84	10	5	leski	a&b) Marasm senilis. Myio insuf	-"- lago	lin

LISARKISTO



B11

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	N:o	Sukunimi	Etunimi	a talouskun	puoli: Mie	puoli: Na	Siviilisäät	yntymävuc	tymäkuuka	ntymäpäir	ntymäkun	mainen asu	ut nykyise	elinkeino k	en työnan	raketta ei	eli: Pääk
2	1	Nieminen	Ale	päämies	1		naimisissa		heinäk¬	19	Teisko	P. Pirkkala	1930	kirvesmie	rak mest.	Fot256	Suomi
3	Tt	Lapsingforsuoja"	Eka	vaimo		1	" "		maalisk	29	Vesilahti	" "	" "	taloustyö	kotona.	0	"
4	III	" palv. Aino Kuokasen	Jaak	poika	1				toukok¬	19	Tampere	" "	"			1	"
5	W	" palouden Wilhelmiir	Mat	" "	1				heinäk¬	22 "	" " 0	" "	"			6	" "
6	T	" Kauppias Alfred Mal	Ant	" palv. Ain	1				helmik¬	13	" O	" "	"			5	" "
7	II	,, Käsityöläin.	Rei	" " 1 "	1.				helmik	13	" O	" "	" "			6	" "
8					5	1											
9																	
10																	
11																	
12																	

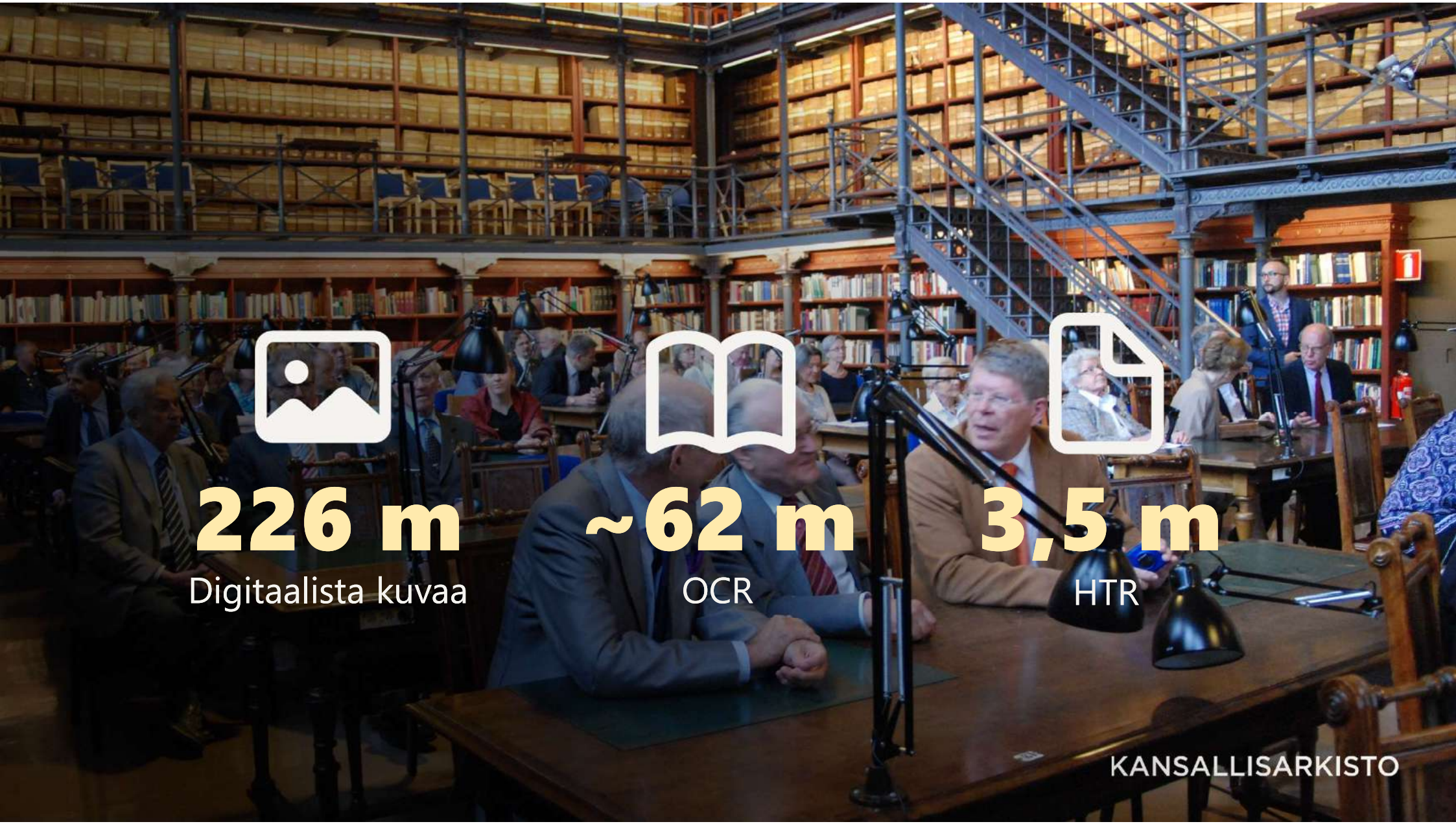

Text: Kauppias MÃ¤kelÃ¤., Score: 0.974
0
0
0
50
100
150
200
250
300
350
400

Tietopyyntöjen tehostaminen

- Tuhansia tietopyyntöjä ajoneuvorekisterin aineistoihin liittyen vuosittain -> vastaaminen vaati paljon työaikaa
- Sisällöntunnistuksella apua tietopyyntötyöhön
  - Kuvia annotoitiin n. 800
  - Yolo segmentointimallina, TrOCR tekstintunnistukseen

- **Tehostaa työtä huomattavasti, arviolta laskee tietopyyntöön vastaamisen kuluvan työajan alle puoleen entisestä**

KANSALLISARKISTO


226 m
Digitaalista kuvaa
~62 m
OCR
3,5 m
HTR
KANSALLISARKISTO

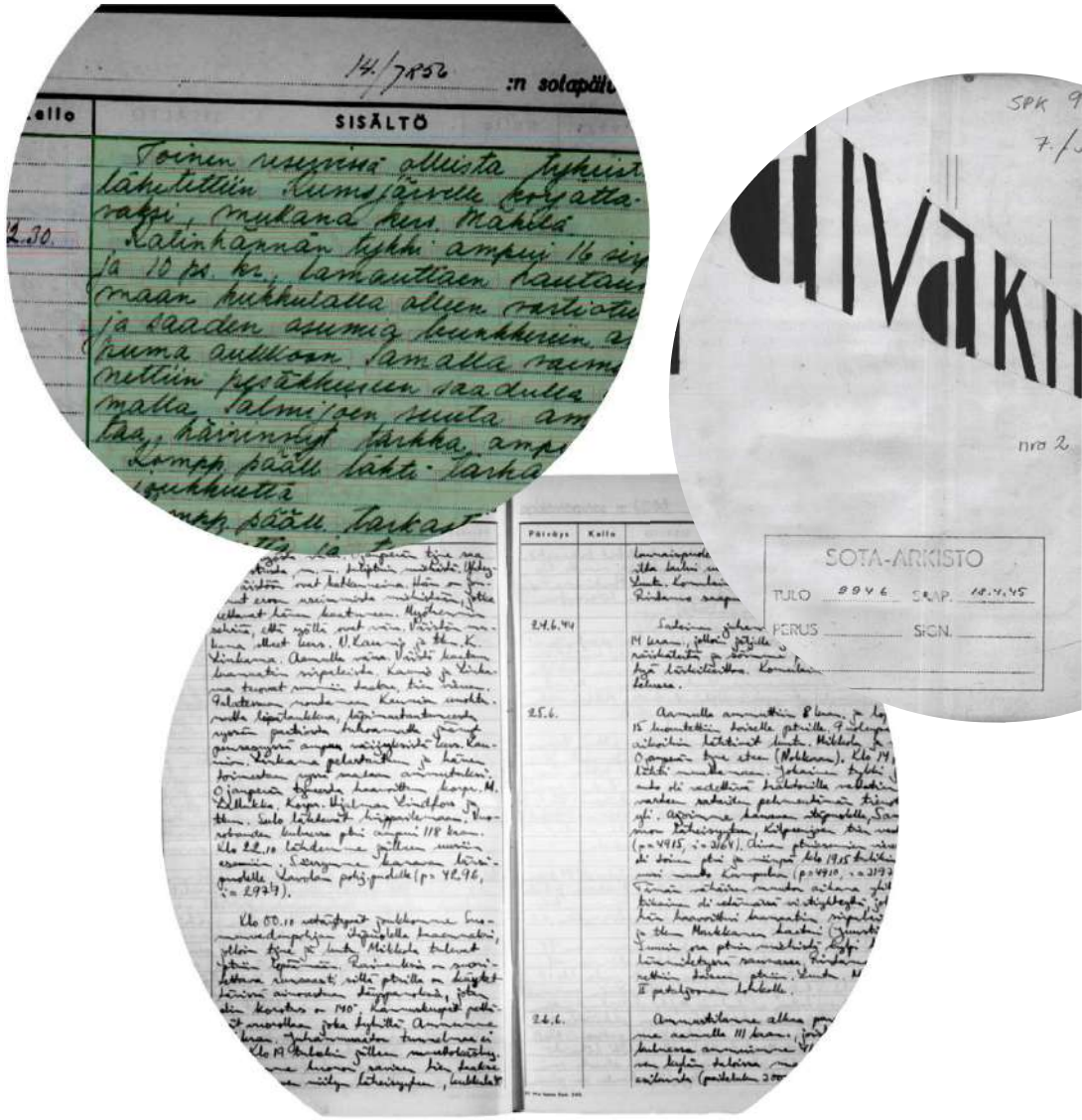
Aineistojen prosessointisuunnitelmia

- KA päässyt mukaan CSC:n akateemisen käytön piiriin
  - Merkittävät laskentaresurssit käyttöön
  - Jatkossa jopa 20 miljoonaa sisältötunnistettua sivua vuosittain
- Parhaillaan prosessoidaan 1800-luvun tuomiokirja-aineistoa
  - Yhteensä hieman alle 10 miljoonaa sivua
  - Käytännössä koko 1800-luvun oikeushistoria sisältöhakujen piiriin vielä tämän vuoden aikana!
- 1800-luvun valmistuttua siirryttään 1700-luvulle ja ensin renovoituihin tuomiokirjoihin

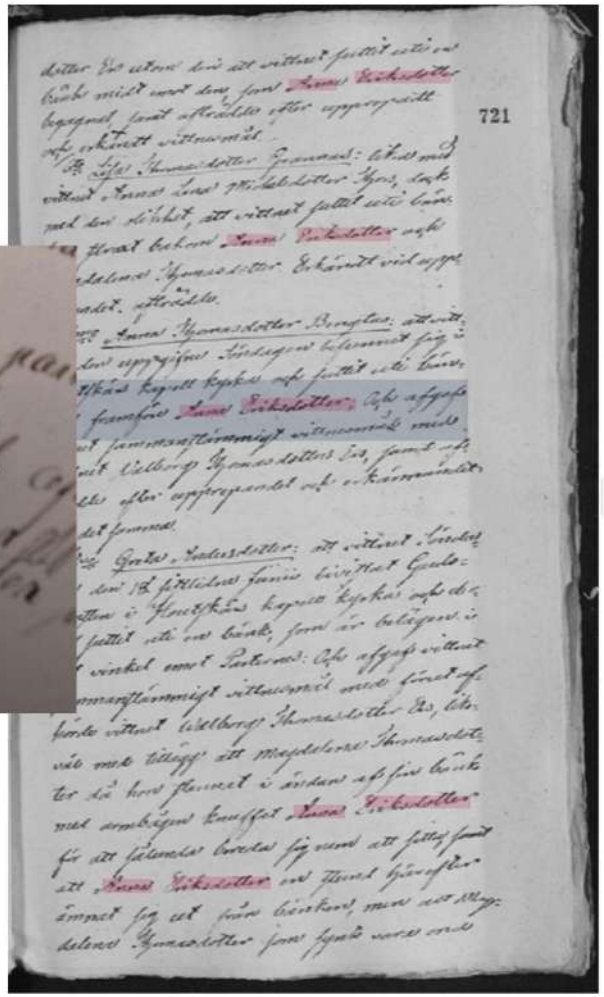
KANSALLISARKISTO

Sotapäiväkirjat

- Kansallisarkiston suurin joukkoistushanke, alkoi keväällä 2022
- Yli 200 vapaaehtoista, joista yli 100 osallistui aktiivisemmin
- Transkribus Lite –työkalun avulla sotapäiväkirjojen puhtaaksikirjoitusta
- Yli miljoona sanaa opetusaineistoa
  - Vaatinut kuitenkin paljon korjailua ja läpikäymistä
- Nyt kuitenkin valmista, ja prosessointi aloitetaan kesään mennessä
  - Taulukkomalli, joka jakaa tapahtumat päivän mukaan



KANSALLISARKISTO



dotter Er utom den att vittnet suttit uti en
bänk midt emot den, som Anna Eriksdotter
begagnat samt afträdde efter upprepadt
och erkändt vittnesmål.
6o Lisa Thomasdotter Grannas: likd med
vittnet Anna Lena Michelsdotter Thors dock
med den denhet att vittnet suttit uti ban¬
ken straxt bakom Anna Eriksdotter och
Magdalena Thomasdotter. Erkändt vid upp¬
repandet. afträdde.
1mo Anna Thomasdotter Bengtas: att vitt¬
net den uppgifne Föndagen befunnit sig i
Houtskän kapell kyrka och suttit uti ban¬
ken framföre Anna Eriksdotter; Och afgaf
vittnet sammanstämmigt vittnesmål med
vittnet Valborg Thomasdotters Ers, samt af¬
trädde efter upprepandet och erkännandet
af det samma.
8° Greta Andersdotter: att vittnet Sönda¬
gen den 18. sistlidne Junii bivistat Gudo¬
tjensten i Houtskäns kapell kyrka och der¬
vid suttet uti en bank, som är belägen i
rat vinkel emot Parternes; Och afgaf vittnet
sammanstämmigt vittnesmål med förut af¬
porde vittnet Walborg Thomasdotter Ers, liks¬

KANSALLISARKISTO

Suomeksi | På svenska | In english | Kontrasti:

Astia

Aineiston pikahaku | Hae

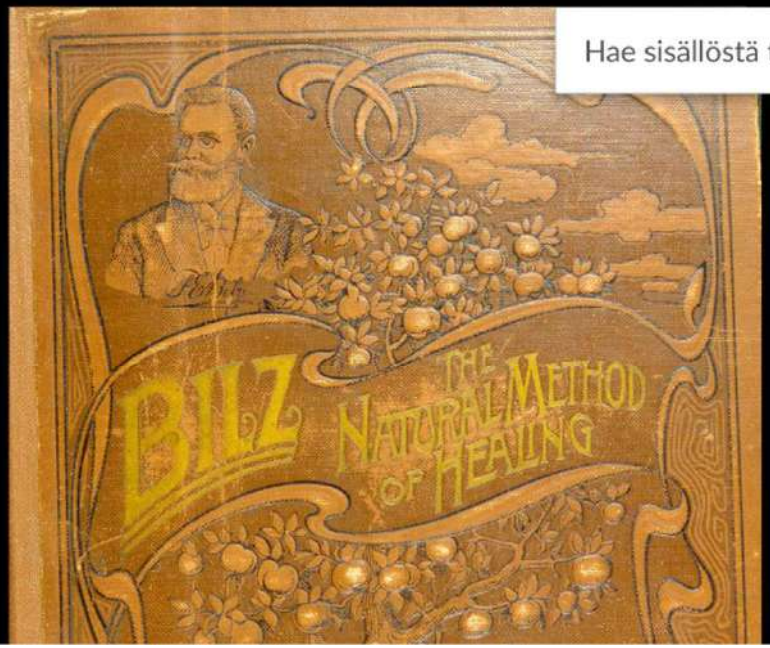
Etusivu | Hae aineistoja | Selaa aineistoja | Tee tietopyyntö | Ohjeet | Palaute

← Takaisin hakutuloksiin | Jaa

The natural method of healing: A new and complete guide to health by F.E. Bilz

Ruudukko | Viitetiedot | Sulje haku ×

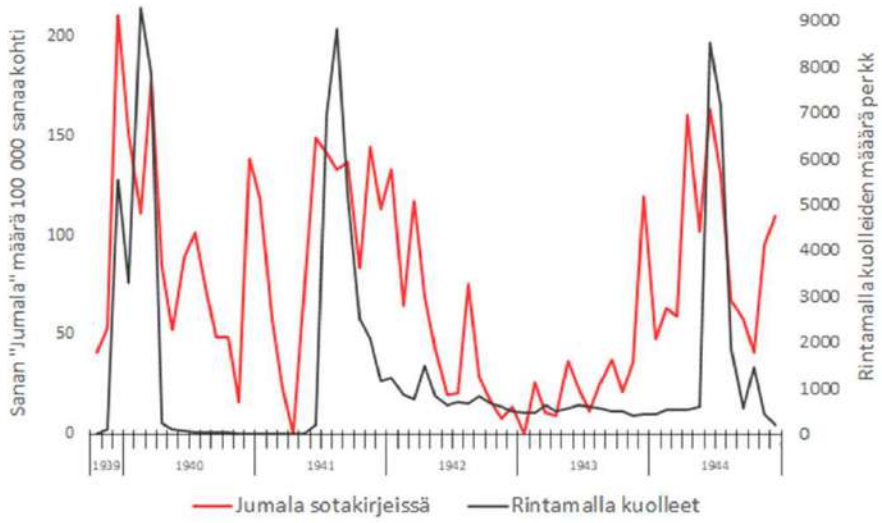
Hae sisällöstä tekstitunnistuksen avulla | Hae

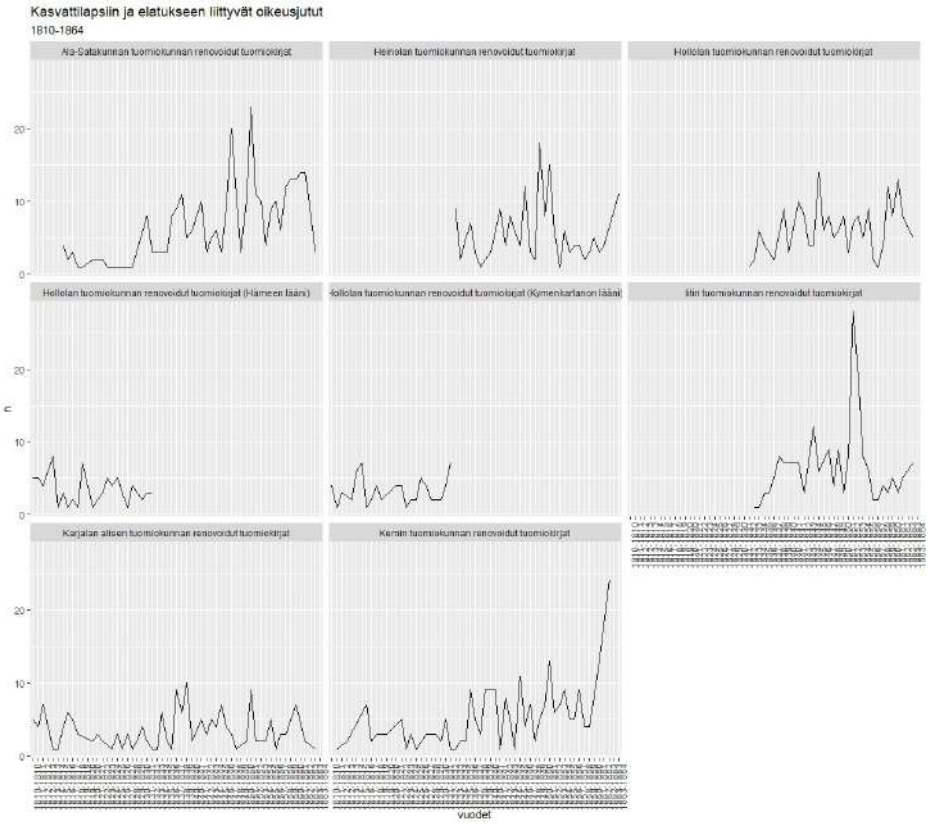


KANSALLISARKISTO

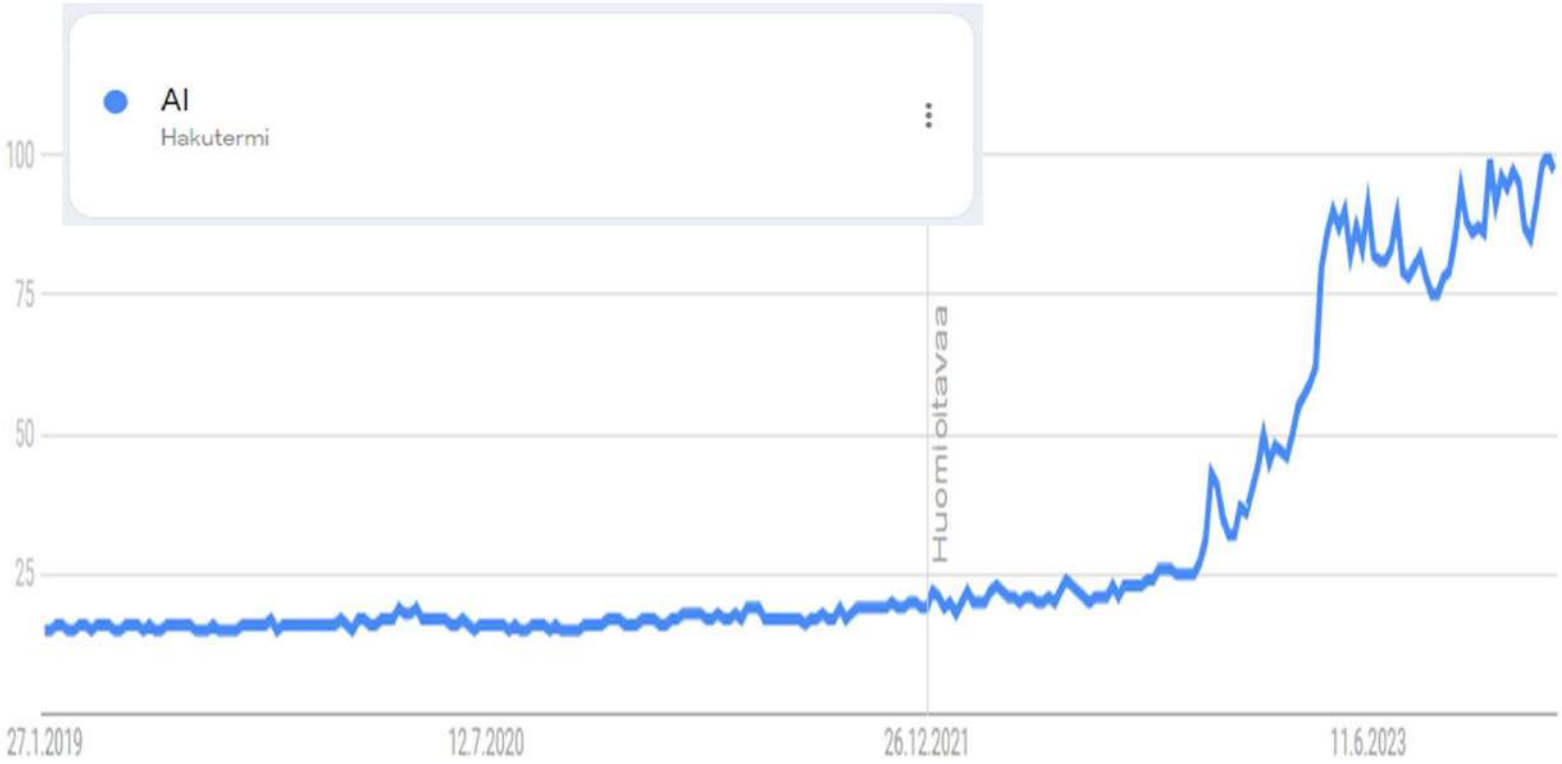
Merkitys tutkimukselle

- Koneluettavuus mahdollistaa uudet menetelmät
- Taulukkomuotoisen aineistot pian mahdollista tuottaa suoraan laskennalliseen tutkimukseen kelpaavaksi dataksi





KANSALLISARKISTO



KANSALLISARKISTO

Kielimallien mahdollisuudet

- Suurten kielimallien potentiaalia asiakirjanhallintaan liittyvissä sovelluksissa ollaan aktiivisesti tutkimassa eri tahoilla
- Mahdollisia käyttökohteita: tiivistelmien teko tekstikokonaisuuksista, kielimallien yhdistäminen asiakirja-aineistoon niin että ne vastaavat käyttäjän aineistoa koskeviin kysymyksiin, kielimallien hyödyntäminen aineistohaussa (esim. vektoritietokannat)…
- Kielimallien käyttöön liittyy kuitenkin monia haasteita:
  - Luotettavuus
  - Maksullisuus
  - Kielivalikoima
  - ‘Raakamallit’ vs. eri käyttötarkoituksiin jalostetut mallit
  - Resurssivaatimukset…
- Parhaillaan kuitenkin kehitteillä avoimia, suomenkielisellä aineistolla koulutettuja malleja
  - TurkuNLP:n GPT-mallit, TukuNLP:n ja Silo AI:n Poro
  - Näiden myötä toivotaan ratkaisuja osaan haasteista

ChatGPT

TurkuNLP / **gpt3-finnish-large**

PORO

KANSALLISARKISTO

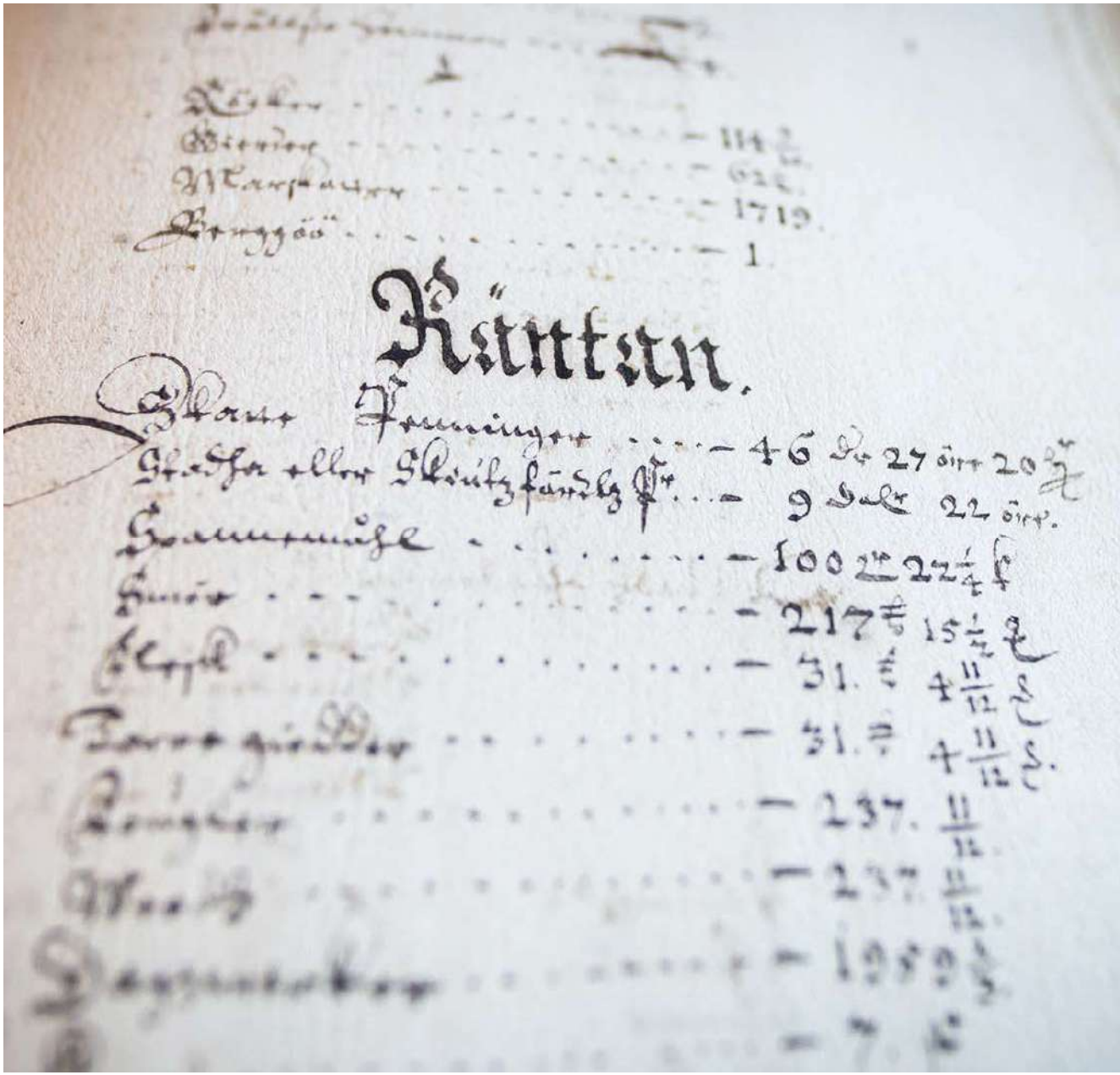
```
, index=index)
bot.load_chat_history("chat_history.json")

while True:
    user_input = input("You: ")
    if user_input.lower() in ["bye", "goodbye"]:
        print("Bot: Goodbye!")
        bot.save_chat_history("chat_history.json")
        break
    response = bot.generate_response(user_input)
    print(f"Bot: {response['content']}")
```

```
You: Milloin suomessa oli sisällissota ja mistä löydän siitä tietoa
Bot:
Suomessa oli sisällissota vuonna 1918. Tietoa sisällissodasta voi löytää Sisällissodan pikkujättiläinen -teoksesta, joka on julkaistu vuonna 2009. Lisäksi Suome
You: Mitä arkistoaineistoja sisällissodasta on?
Bot:
Vapaussodan arkisto sisältää hyvin laajasti aineistoa kansalaissodasta, mukaan lukien punaisen ja valkoisen puolen organisaatioiden tuottamia asiakirjoja, yksit
You: Entä muita?
Bot:
Muita lähteitä, joista voi etsiä tietoa valkoisen armeijan puolella toimineista henkilöistä, ovat kunniamerkkivaliokunnan arkisto ja erilaiset valkoisen puolen
You: Mistä löydän tietoa punaisista?
Bot:
Vapaussodan arkistosta voi löytää hyvin monenlaista tietoa niin sotilaallisista kuin siviilipuolen asioista. Aineisto on jaettu arkistossa keräyspiirin mukaan j

Vapaussodan arkistoa voi tutkia Kansallisarkiston Helsingin toimipisteessä. Sieltä löytyvät myö
You: Mistä löydän tietoa sisällissodassa taistelleista punaisista?
Bot:
Tietoa sisällissodassa taistelleista punaisista voi etsiä Suomen Vapaussota vuonna 1918 -teoksen ja Suomen Vapaussota -teoksen lisäksi muista lähteistä, kuten s
You: Mistä voin löytää tietoa yksittäisestä punaisten puolella taistelleesta henkilöstä?
Bot:
Tietoa yksittäisestä punaisten puolella taistelleesta henkilöstä voi etsiä punaisten puolella taistelleiden joukko-osastojen arkistoista. Kansallisarkistossa He
You:
```

KANSALLISARKISTO

Kiitos!

Ilkka.jokipii@kansallisarkisto.fi
mikko.lipsanen@kansallisarkisto.fi



Esitys on toteutettu osana Kaakkois-Suomen ammattikorkeakoulu Xamk:n, Ammattiopisto Samiedun ja Kansallisarkiston yhteishanketta JoDi-Joustavat koulutus ja työelämäpolut tulevaisuuden digitointiosaajille (ESR+ 2023-2025).

SAMiEDU
KANSALLISARKISTO
Elinkeino-, liikenne- ja ympäristökeskus
Euroopan unionin osarahoittama
XAMK

KANSALLISARKISTO

www.kansallisarkisto.fi

@kansallisarkisto

@kansallisarkist