



KANSALLISARKISTO

# Digitoidun aineiston sisällöntunnistus Kansallisarkistossa

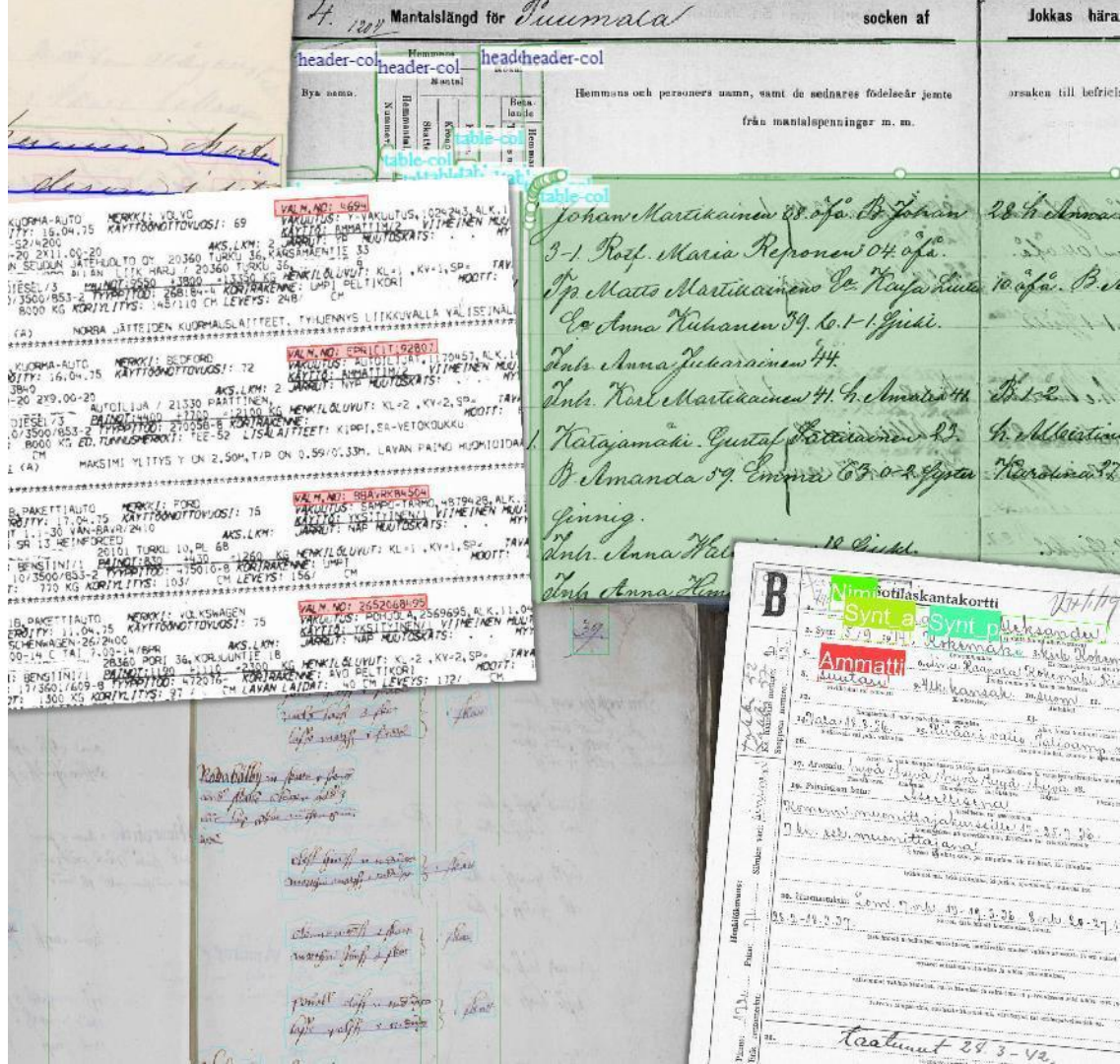
Viranomaisinfo 12.3.2026

Sanna Joska, tutkija

Data, tekoäly ja sovelluskehitys

[viranomaisohjaus@kansallisarkisto.fi](mailto:viranomaisohjaus@kansallisarkisto.fi)

13.3.2026



# Digitoidun aineiston sisällöntunnistusta jo vuodesta 2016

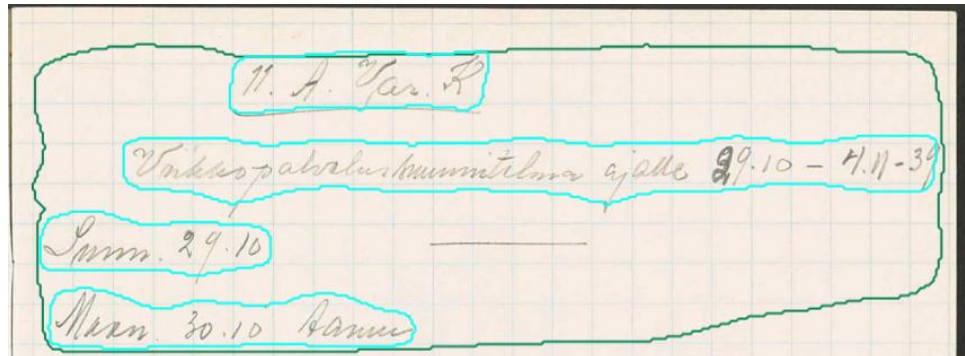
Kansallisarkisto on  
kehittänyt digitoidun aineiston  
sisällöntunnistusta jo vuodesta 2016

- Koneoppimisen asiantuntijat
- Tuotekehitys
- Hankkeet
- 2026: Tekoälyosaamisen laaja-  
alainen hyödyntäminen  
organisaatiossa



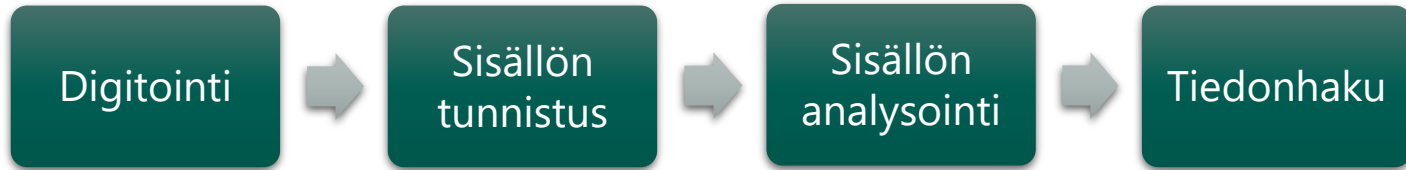
# Sisällöntunnistus

- Sisällöntunnistus on automaattinen prosessi, jossa digitaalisesta aineistosta tunnistetaan ja luokitellaan informaatiota
  - Teksti, kuvat, rakenteet, metatiedot
- Tekoälyyn ja koneoppimisteknologiaan (syväoppiminen) perustuvat mallit
- Laadukas opetusaineisto takaa usein hyvät sisällöntunnistustulokset
- Sisällöntunnistus parantaa olennaisesti suurten aineistomäärien käytettävyyttä
  - Mahdollistaa esim. haut, indeksoinnin ja muun datan jatkokehityksen



- 1 11. A. Var. K.
- 2 Viikkopalvelussuunnitelma ajalle 29.10 - 4.11-39
- 3 Sunn. 29.10
- 4 Maan. 30.10 Aamu

# Tekoäly asiakirjahallinnan eri vaiheissa



# Kansallisarkiston sisällöntunnistusmallit 1/2

## 1500–1900-lukujen käsialamalli (HTR-malli)

- Jatkokoulutusaineistona lähes miljoona riviä 1500–1900-lukujen käsinkirjoitettua suomen- ja ruotsinkielistä tekstiä
- 1700–1800-luku vahvimmin edustettuna aineistossa, näiden osalta tunnistustulokset myös parhaita
- Merkkivirhe keskimäärin n. 3%
- Koulutusaineistoa pyritään lisäämään ja mallia päivittämään mahdollisuuksien mukaan – esim. viimeisimpään (syksy 2025) malliin otettiin mukaan myös 1500-luvun aineistoa

## Konekirjoitetun tekstin tunnistusmalli (OCR-malli)

- Jatkokoulutettu yli 150 000 rivikuvan aineistolla (pääosin suomen- ja ruotsinkielistä)
- Konekirjoitetulla aineistolla merkkivirheprosentti 1.24–2.29%

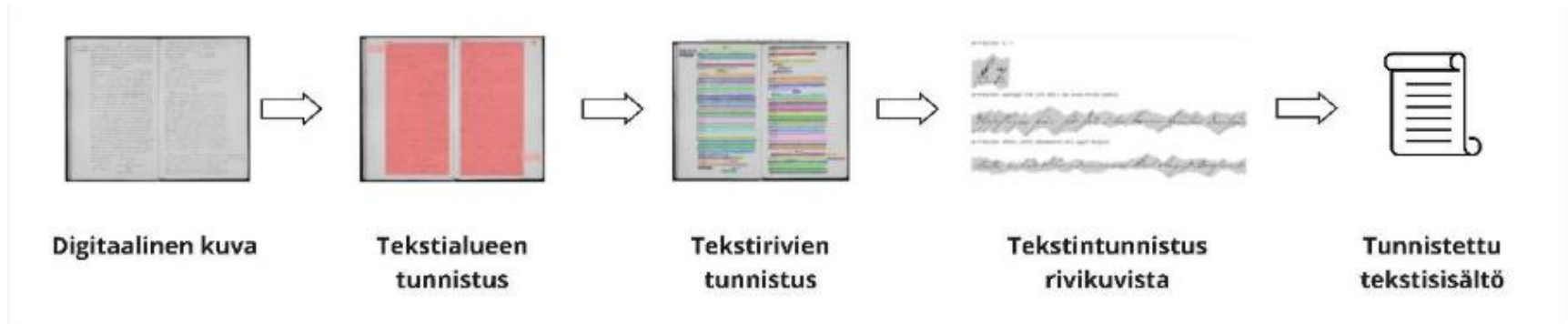
## Kansallisarkiston sisällöntunnistusmallit 2/2

- **Segmentointimalli** eli tekstialue- ja rivimalli tunnistaa rakenteet kuvasta
  - Kehitetään sopimaan yhä useammille aineistotyypeille
- **Kyrillinen käsialamalli:** 1800–1900-lukujen venäjänkielistä käsialaa osaava malli kehitteillä EU:n Interreg Central Baltic –rahoitteisessa ArchXAI-projektissa
- **NER-malli** (DALAI-hanke) tunnistaa aineistosta nimientiteettejä, kuten henkilöiden ja organisaatioiden nimiä, y-tunnuksia, päivämääriä etc.

Valtion Lisenssitoimikunnalle **ORG** . Saksalainen **NORP** kirjailija, tri Friedrich Wolf **PERSON** , Berlin **GPE** - **PERSON** Pankow **GPE** , on Suomen Yleisradiolle **ORG** luovuttanut 2 käsikirjoituksensa esitysoikeuden, joista käsikirjoituksista toinen on jo esitetty ohjelmistossa ja toinen tullaan lähiaikoina esittämään. Esitysoikeuden luovuttamisesta on tri Wolf **PERSON** saamassa yhtiöltämme pakkionsa, jonka hän meille saapuneen tiedon mukaan toivoisi saavansa vastaanottaa elintarvikkeiden muodossa. Tämän vuoksi Oy.Yleisradio Ab. **ORG** anoo kunnioittavasti, että Valtion lisenssitoimikunta **ORG** myöntäisi yhtiöllemme vientilisenssin 10 kg:lle lihaa, sen lähettämistä varten tri Wolfille **PERSON** . Lisenssit pyydämme saada kahta 5 kg:n lähetystä varten. Helsingissä **GPE** , tammikuun 20 p:nä 1948 **DATE** . Hella Wuolijoki **PERSON** . Einar Sundström **PERSON** .

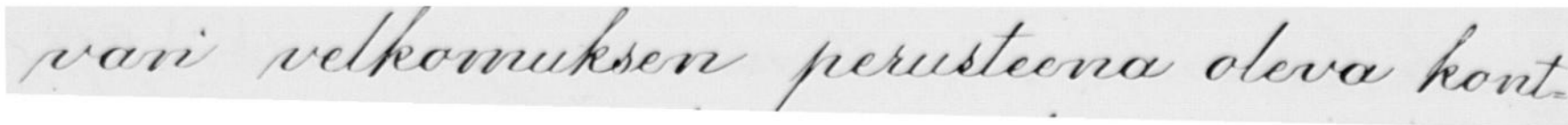


# Käsialantunnistusprosessi



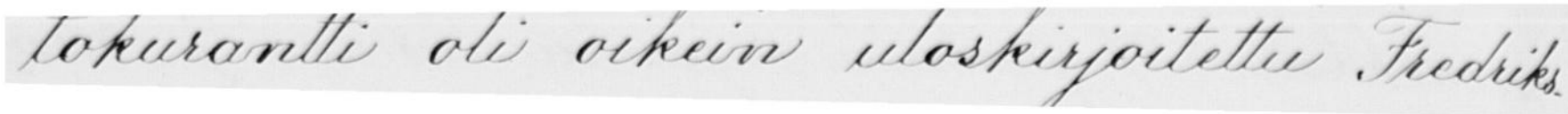
# Sisällöntunnistus: käsinkirjoitetun tekstin tunnistus (HTR) 1/2

4509# Predicted: van velkomuksen perusteena oleva kont-  
Original: van velkomuksen perusteena oleva kont-



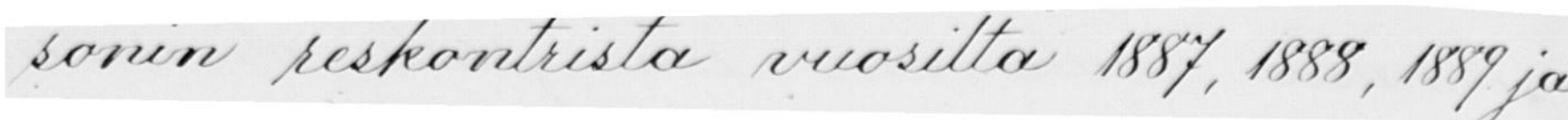
van velkomuksen perusteena oleva kont-

4510# Predicted: tokurantti oli oikein uloskirjoitettu Fredriksi  
Original: tokurantti oli oikein uloskirjoitettu Fredriks-



tokurantti oli oikein uloskirjoitettu Fredriks

4511# Predicted: sonin reskontrista vuosilta 1887, 1888, 1889 ja  
Original: sonin reskontrista vuosilta 1887, 1888, 1889 ja



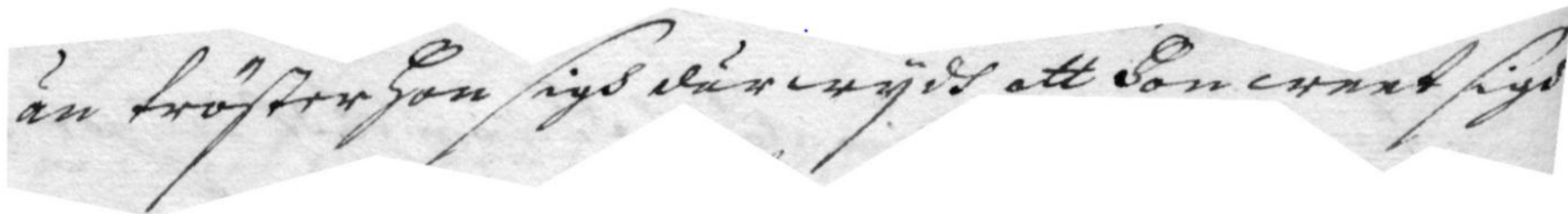
sonin reskontrista vuosilta 1887, 1888, 1889 ja



# Sisällöntunnistus: käsinkirjoitetun tekstin tunnistus (HTR) 2/2

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

Original: än tröster Hon sigh därwydh att Hon weet sigh

A snippet of handwritten text on aged paper, showing the words 'än tröster hon sigh därwydh att hon weet sigh' in a cursive script.

84# Predicted: wara oskyllldigh, och tackar Gudh därföre

Original: wara oskyldigh, och tackar Gudh därföre

A snippet of handwritten text on aged paper, showing the words 'wara oskyldigh, och tackar Gudh därföre' in a cursive script.



Kuolleet

Haminan

seurakunnassa v:n 1936

neljänneksellä.

(Lähetettävä Tilastolliseen päätoimistoon 15 päivän kuluessa neljänneksen päättymisestä)

| Kuoleman aika | Sivu<br>pää- | Kuolleen nimi, sääty ja osoite | Syntymä-<br>aika | Ikä | Naimen,<br>Leski, Br<br>Tapp | Kuoleman<br>syy | Miehen | Vainon |
|---------------|--------------|--------------------------------|------------------|-----|------------------------------|-----------------|--------|--------|
|               |              |                                |                  | V   | M                            | P               |        |        |

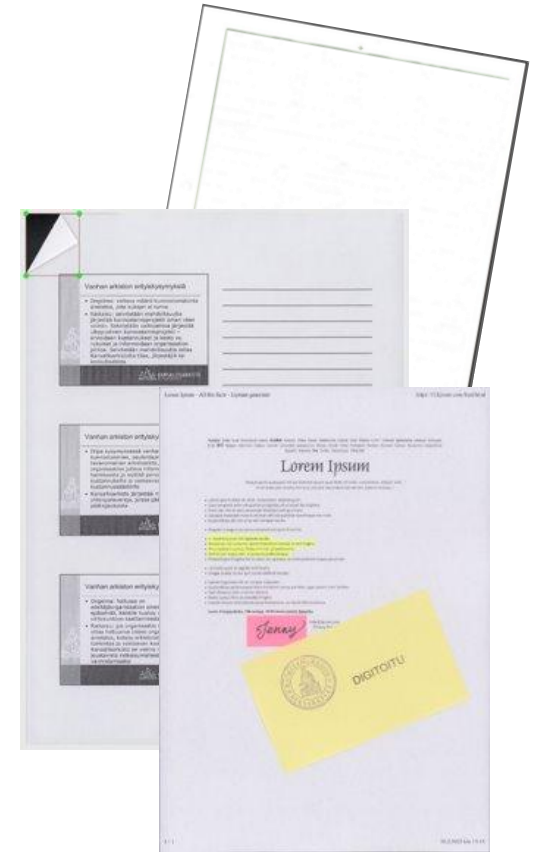
|    | A        | B     | C          | D                                     | E            | F           | G       | H           | I          | J                                                     | K           | L        |
|----|----------|-------|------------|---------------------------------------|--------------|-------------|---------|-------------|------------|-------------------------------------------------------|-------------|----------|
| 1  | Kuukausi | Päivä | u pääkirja | Kuolleen nimi, sääty ja osoite        | Syntymäaika  | Ikä: Vuotta | Kuukaut | Ikä: Päivää | ut, Leski, | Kuoleman syy                                          | Miehenp.    | Vaimonp. |
| 2  | Tammik.  | 12    | 591        | Saikkonen, Iida Maria, työm. vo       | 7.I 1884     | 52          |         | 5           | Nainut     | a) Carcinoma cordis e anemia secundis                 |             | 1        |
| 3  | "        | 15    | 546        | Rautaleins, Matilda, johtajan leski   | 13.III. 1876 | 59          | 10      | 2           | leski      | a) Myodegeneratio cordis (3030 sar                    | 81          | lis      |
| 4  | "        | 16    | 920        | Lautala, Antti, kalastaja             | 22.XII       | 95          |         | 24          | "          | a) Haama- raspiacordis                                | 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) Tuberculum pulmo. b) sa                            | s. mini     | 1        |
| 8  | Helmik.  | 5     | 512        | Pousi, Anna Maria, leski              | 20.I. 1867   | 68          |         | 15          | leski      | a) Myodegeneratio insufficiens cordis atio tio siv    | 1           |          |
| 9  | "        | 13    | 689        | Takalahti, Kaarle Antinp. talonmies.  | 25.I. 1865   | 71          |         | 18          | nton       | a) Myodegeneratio et insufficia cordia sa 7/5 ole in. |             |          |
| 10 | "        | 17    | 44         | Betán, Mariaana, leski                | 20.V 1874    | 61          | 8       | 27          | leski      | a) Vitium ca 3020; b) Insufficiencia cor              | Pelin       | 1        |
| 11 | "        | 13    | 71         | Eriksson, Aron Johan, ent. pol. konst | 15.VI 1876   | 59          | 7       | 28          | n-nut.     | a) Hypertonia arterialis. mon                         | 1           |          |
| 12 | "        | 25    | 290        | Kitunen Helena. leski                 | 6.1. 1849    | 87          | 1       | 20          | leski      | a) Marasmus senilis; b.) Bronchum velko-              |             | 1        |
| 13 | Maalisk  | 19    | 426        | Naakka, Amanda, työmiehen vaimo       | 6.9.1898     | 37          | 6       | 13          | n-nut      | a) Tuberculo meningium                                | sia. bää in | 1        |
| 14 | "        | 23    | 73         | Eriksson, Vilhelm. työmies            | 6.8 1862     | 73          | 7       | 17          | "          | a) Samoon hepatio. b) samoin.                         | 1           |          |
| 15 | "        | 28    | 464        | Paavola, Maria Kristiina, leski       | 23.5. 1851   | 84          | 10      | 5           | leski      | a&b) Marasmus senilis. Myo insuf                      | -"- lago    | lin      |

|         |    |     |                                 |            |    |    |    |       |                                               |   |  |
|---------|----|-----|---------------------------------|------------|----|----|----|-------|-----------------------------------------------|---|--|
| "       | 25 | 290 | Kitunen, Helena, leski          | 6.1. 1849  | 87 | 1  | 20 | leski | a) Tuberculo meningium                        | 1 |  |
| Maalisk | 19 | 426 | Naakka, Amanda, työmiehen vaimo | 6.9. 1898  | 37 | 6  | 13 | n-nut | a) Tuberculo meningium et insufficiens cordis | 1 |  |
| "       | 23 | 73  | Eriksson, Vilhelm, työmies      | 6.8 1862   | 73 | 7  | 17 | "     | a) Samoon hepatio. b) samoin.                 | 1 |  |
| "       | 28 | 464 | Paavola, Maria Kristiina, leski | 23.5. 1851 | 84 | 10 | 5  | leski | a) Marasmus senilis. Myo insuf                | 1 |  |

LISARKISTO

# Tulossa: Analytiikkainfrastruktuuri osana massadigitointia

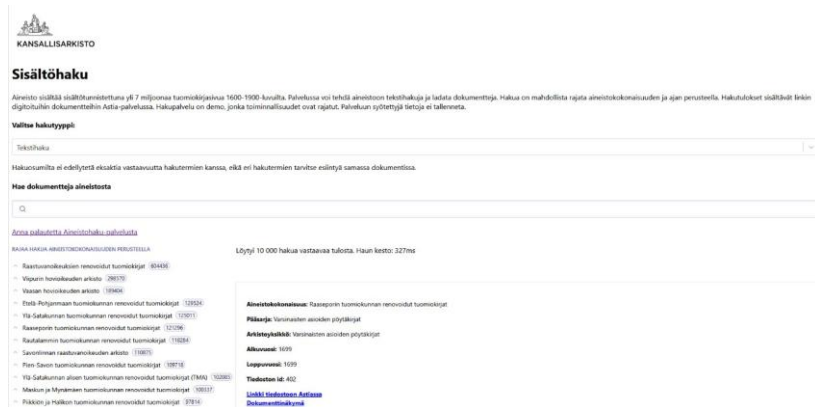
- Analytiikkainfrastruktuuri tarjoaa sisällöntunnistuspalvelut osaksi massadigitoinnin prosessia
- Tekstintunnistus (HTR ja OCR)
- Digitoinnin automaattisen laadunvarmistuksen komponentit (DALAI-hanke). Jos virhe havaitaan, aineisto skannataan uudestaan:
  - tyhjien sivujen tunnistus
  - post-it-lappujen tunnistus
  - taittuneiden kulmien tunnistus
  - Rotaation tunnistus
- Viranomaisten tarpeista nouseva sisällöntunnistuksen kehitystyö, esim. tunnistustuloksen laadunparannus



# Tiedonhaku sisältötunnistetuista aineistoista

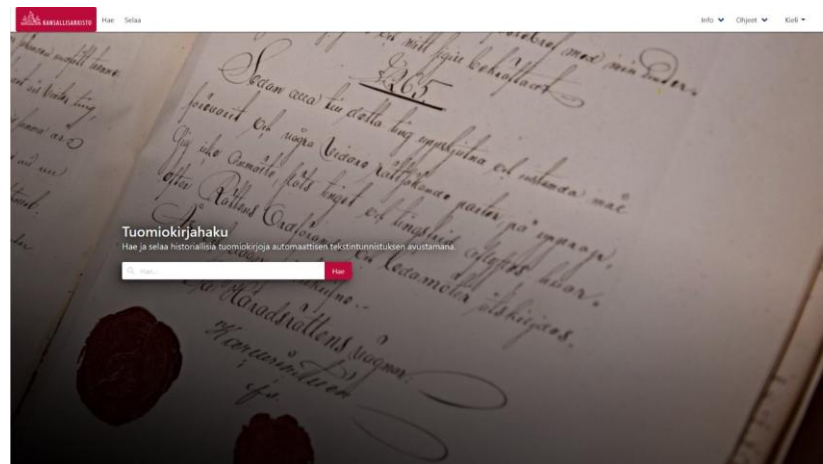
## Sisältöhaku

<https://sisaltohaku.demo.kansallisarkisto.fi/>



## Tuomiokirjahaku

<https://tuomiokirjat.kansallisarkisto.fi/#/>



# Tiedonhaku: MCP



**MCP (Model Context Protocol) on avoimen lähdekoodin protokolla, joka yhdistää tekoälymallit ulkoisiin tietolähteisiin ja palveluihin** – sen sijaan että tekoäly olisi rajoitettu vain harjoitusdataansa, se voi reaaliaikaisesti hakea tietoa tietokannoista, API:sta ja paikallisista järjestelmistä.

- Kansallisarkisto tehnyt kokeiluja MCP-protokollan kanssa
- Tekoälypalvelun (esim. ChatGPT, Claude) kautta mahdollista tehdä hakuja palveluihin luonnollisella kielellä

SJ

Etsin tietoa lapsen elatusasioista Turun ja Porin läänin alueelta 1900-luvun alusta. Etsi, paljonko lapsen elatusasioita on käsitelty käräjillä ja kerro pääpiirteitä jutuista.

Etsin Turun ja Porin läänin alueen tuomiokirjoista tietoa lapsen elatusasioista 1900-luvun alusta.

N

Haun rajaaminen ajan ja paikan mukaan

N

Tekstihaku

Etsin nyt tarkemmin nimenomaan Turun ja Porin lääniin kuuluvia tuomiokirjoja sekä hakusanalla "lapsen elatus":

N

Tekstihaku

Request

```
{
  `size`: 30,
  `text`: `lapsen elatus`,
  `index`: `tuomiokirjat`
}
```

Response

```
[{"text": "Total results: 490, showing 30 hits from position 0"},
```

Erinomaista! Löysin paljon aineistoa. Nyt etsin tarkemmin Turun ja Porin lääniin kuuluvia tuomiokirjoja. Etsin erityisesti Porin raastuvanoikeuden ja muiden Turun ja Porin läänin alueen tuomiokuntien dokumentteja:

N

Haun rajaaminen ajan ja paikan mukaan

Loistava! Löysin 21 tapausta Porin raastuvanoikeudesta. Nyt etsin lisää tietoa Turun seudulta:







# KANSALLISARKISTO



[kansallisarkisto.fi](https://kansallisarkisto.fi)



@kansallisarkisto



@kansallisarkisto.fi