



FIN-CLARIAH 2024/25

D:3.3.1 Enhancing the usability of archival data

Authors: Antero Holmila University of Jyväskylä Ida Toivanen University of Jyväskylä Venla Poso University of Jyväskylä Tanja Välisalo The National Archives of Finland Mikko Lipsanen The National Archives of Finland

Introduction

The National Archives of Finland is digitising documents across a range of fields, encompassing both historical archives and more contemporary state authority records. Our work focuses on enhancing the accessibility and usability of these documents through the application of machine learning techniques. By automatically recognizing content from unstructured digitized data, we unravel various potential applications for archival data.

Enriching the archival documents

In order to make digitised archives useful for different user groups, the unstructured archival data needs to be enriched in different ways. Recognising content and creating metadata can help to organise, search and analyse the data. For example, recognising journal numbers (*diaarinumero*) can be used to correctly date the documents.

Our research aims to explore both the visual and textual features of digitized archival data. We have developed a named entity recognition (NER) model designed to identify significant entities within archival data. Additionally, we are working on creating a solutions for automatically recognizing various table structures and their contents. This work will build on previously done development with table recognition and HTR in the NAF. The ultimate objective is to benefit end-users in different fields of research and administration by enhancing the usability of archival data.

Archival data

The inherent challenge of archival data is in its heterogeneous nature. The archival data originates from diverse sources and spans various historical periods. The temporal aspect introduces complexities not only due to the evolution of formats and writing styles but also because digitisation practices and image quality vary across documents produced in different eras.



Task 1: Named entity recognition

In the NER model development we concentrated on two main challenges: 1) on reducing the effect of OCR noise and 2) idiosyncrasies of formal language in mass digitised state authority data. First version of named entity recognition model developed for the archival data is available in Huggingface, but we will publish the final model in the fall with annotation guidelines.



huggingface.co/ <u>Kansallisarkisto</u> <u>/finbert-ner</u>

ulinder - - - - -

Task 2: Table recognition

Starting with testing existing HTR and table recognition models and then focusing on researching the possibilities of **multimodal approaches** to the archival data, we aim to find sustainable solutions for recognising tables and their content from both historical and modern archival data.

6. Voglan hede Nygard. Fødelse: Kommen 3 uner minnet: 7 955	Nattvardsgång.		6.	. Yalan hide Mygard.
1833. 1833. 1833. 1833. 1833. 1833. 1833. 1833. 1833. 1833. 1833. 1833. 1833. 1833. 1833. 1833. 1834. 1944. 19	1834 1835 1836 1837 1838 1839 Anmärkningar. Argin.			Födelse: Komman Komm
I marin Simply Rate 21 21 21 24 24 24 24 24 24 24 24 24 24 24 24 24	L ILMOITUS LAPSEN OTTAMISEKSI VÄESTÖREKISTERIIN		27 ST	Stoken Abraham Ball
Son Carl Gugher 5:36 2	A Osoite ISA Henkilötunnus B Vihkimispaiva Väestörekisteri (jollei sama kuin äidin) Osoite (jollei sama kuin äidin)	"A": { "ÄITI": {	{	the Marrie Standites Har 23 . 21
A Sola Territe Specifier 18\$17 Weber 200 W XX XX XX XX & & T 5.45 X Saffress of Unix all 18200 Large Dynam S & 10 XX XX Saffress of Unix all 18200 Large Dynam S & 10 XX XX Saffress of Unix all 18200 Large Dynam S & 10 XX XX Saffress of Unix all 18200 Large Dynam S & 10 XX XX XX Saffress of Unix all 18200 Large Dynam S & 10 XX XX XX Saffress of Unix all 18200 Large Dynam S & 10 XX XX XX Saffress of Unix all 18200 Large Dynam S & 10 XX XX XX Saffress of Unix all 18200 Large Dynam S & 10 XX XX XX Saffress of Unix all 18200 Large Dynam S & 10 XX XX XX Saffress of Unix all 18200 Large Dynam S & 10 XX XX XX Saffress of Unix all 18200 Large Dynam S & 10 XX XX XX Saffress of Unix all 18200 Large Dynam S & 10 XX XX XX Saffress of Unix all 18200 Large Dynam S & 10 XX XX XX Saffress of Unix all 18200 Large Dynam S & 10 XX XX XX Saffress of Unix all 18200 Large Dynam S & 10 XX XX XX Saffress of Unix all 18200 Large Dynam S & 10 XX XX XX Saffress of Unix all 18200 Large Dynam S & 10 XX XX XX Saffress of Unix all 18200 Large Dynam S & 10 XX XX XX Saffress of Unix All 18200 Large Dynam S & 10 XX XX XX Saffress of Unix All 18200 Large Dynam S & 10 XX XX XX Saffress of Unix All 18200 Large Dynam S & 10 XX XX XX Saffress of Unix All 18200 Large Dynam S & 10 XX XX XX Saffress of Unix All 18200 Large Dynam S & 10 XX XX XX Saffress of Unix All 18200 Large Dynam S & 10 XX XX XX Saffress of Unix All 18200 Large Dynam S & 10 XX XX XX XX Saffress of Unix All 18200 Large Dynam S & 10 XX XX XX XX XX Saffress of Unix All 18200 Large Dynam S & 10 XX	APSI Syntymapäivä tai henkilötunnus Sukupuoli Järjestyskirjain 1 poika 2 tyttö O avioliitossa syntynyt 1 avioliiton ulkopuolella syntynyt Sumer kansalainen Muu kansalaisuus	"Henkile" "value"	lötunnus": { .": "",	The Helman Contrator 184216 Setter States W XX XX XX Su. Im Carl Cultor Brost X
× Song Chargen Walling 18513 - Then I adverted to Brot to 4 × Song Them The to find 18500 - I adverted to Brot to 4 × To main them Carlo I 18010 Check park 2 doctor and Su G to C × Gift the field coment Juge 11800 m 20 June 5 40 40 + 5 Bi G 4	Syntymäkottiuutta Uikomailla syntymeen syntymavaltio Syntymäkottiuutta Lisatietoja Paivays	"coord: },	linates": [110, 80, 400, 100]	Saturda Contrato Prover marte 11. 3. 27 9. 20 20 20 20 20 20 20 20 20 20 20 20 20
* He Canadamile mant star 12 be 33 St sarres in Par 4 + To tamel Seman 1812 + Corpe 8° B.	Leima Vaestorekisterin oitajan allekirjoitus tai nimil LAPSEN HUOLTAJAT TÄYTTÄVÄT D-rivikön (Ohjeita alimmaisen kappaleen kääntöpuolella)	"Sukuni "value"	imi": { ;": "",	
- Clasphing Va 20 200 Tite 21	Lapselle vaestokirjoihin merkittavat etunimet Lapselle vaestokirjoihin merkittavat etunimet Lapselle vaestokirjoihin merkittavat etunimet 1 suomi 2 ruotsi Tiedot antoivat (paivavs ja allekirjoitukset)	"coord: },	linates": [420, 80, 800, 100]	a the man hand hand the said of a creat of the
File Same Vitant 182 39	KASTEILMOITUS EVLUT. KIRKON TAI ORTODOKSISEN KIRKKOKUNNAN SEURAKUNNAN JÄSENEKSI REKISTERÖITÄVASTÄ LAPSESTA Sukunimi ja erunimet	"Etunim" "value"	net": { ;": "",	1 2 4 2
Images Confirmation	Onko lapsi hatakastettu Kastepaikka Kastepaiva ja kastajan allekirjoitus kyllä ei LAPSEN OTTAMISTA MUUN USKONTOKUNNAN KUIN EVLUT, KIRKON TAI ORTODOKSISEN KIRKKOKUNNAN SEURAKUNNAN JÄSENEKSI KOSKEVAT TIEDOT Lapsi olettu jaseneka, pym Uskonnolisen yridyskunta seurakunta	"coord" },	linates": [110, 110, 800, 130]	V to at a mining with the and the second of the
catalogue (1833-1839) and	Lapsen REKISTERÖINTIA VÄESTÖREKISTERIIN KOSKEVAT TIEDOT Rekisteröintipaiva Vaestorekisteri, jonka kirjoihin otettu Lasna/poissaoleva 1 LV 2 PV Lapselle vaestökirjoihin merkitty sukunimi Kunta, jonka poissaolevaan vaestöön merkitty	"Usoite" "value"		2 PLLL PLLL PLIST VILLE & U
Notification of a child's	I Lauseile vaestokirjoinin merkityt etunimet İsyyden vahvistamisbaiva	coord	mates: [110, 140, 800, 100]	A Maria 18- st. 4.

18 2 at.

Notification of a child's personal details form.

	`	1.1.1.1.1.1.1		distribution of the second s		o frittriaden idetteloir ino	
1 ei		2 kvllä	Pvm		ei		
				Paivays			
	Leima			Vaestorekisterin eitajan allekirjoitus tai nimileima			



Collaboration

This work will be conducted in cooperation with the National Archives of Finland and the University of Jyväsylä.

