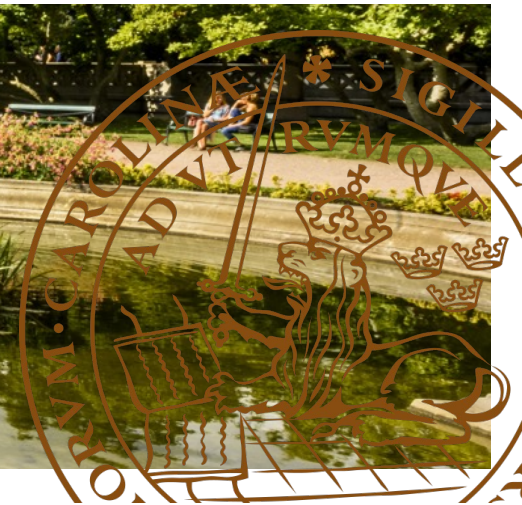




LUND
UNIVERSITY

Pilotprojektet porträttkatalogen

ÅSA FORSBERG OCH MARIA HEDBERG, UB



Arbetsuppgifter i pilotprojektet

- Bildfångst (bilderna i porträttkatalogen var redan inscannade)
- Segmentering (hitta var på bilderna det finns text)
- Transkribering
- Lagring och indexering av metadata
- Rättningar av transkriberingen
- Rättningar av metadata
- Visning av katalog



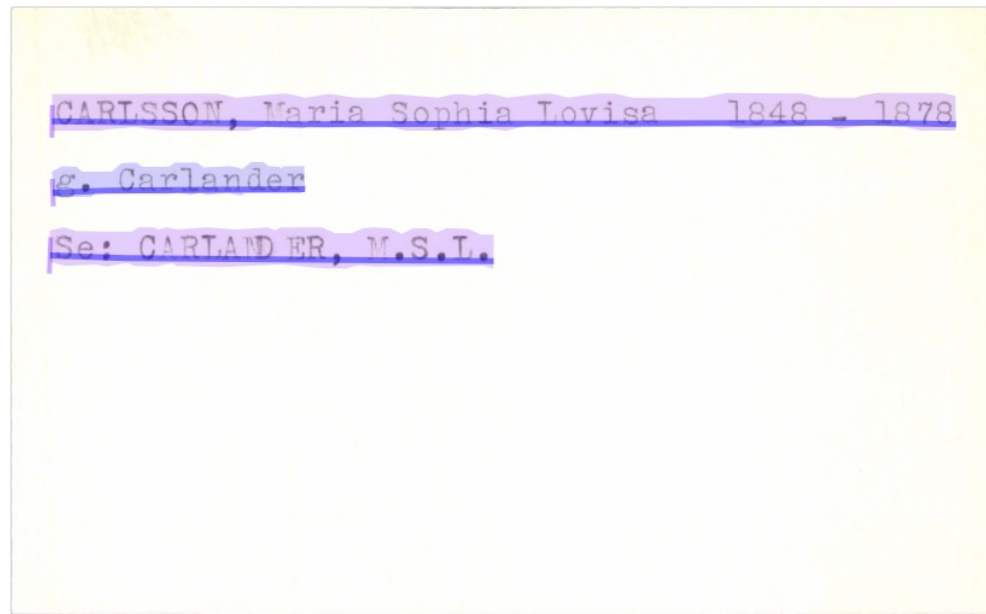


Målet med kortkatalogprojektet är att de objekt som är registrerade i våra kortkataloger ska bli "hittbara" online och därmed tillgängliga för våra användare

Målet med pilotprojektet är dessutom att testa olika lösningar, få erfarenheter och lära oss hur vi ska hantera denna och de andra kortkatalogerna

Segmentering

- För att kunna transkribera texten behöver man först hitta var texten finns på bilden
- I Transkribus finns en fungerade segmentering men den går inte att träna
- Vi har skapat egna segmenteringsmodeller med AI-verktyget Kraken som går att träna, och därmed förbättra
- Vi har testat rättning av segmentering i eScriptorium

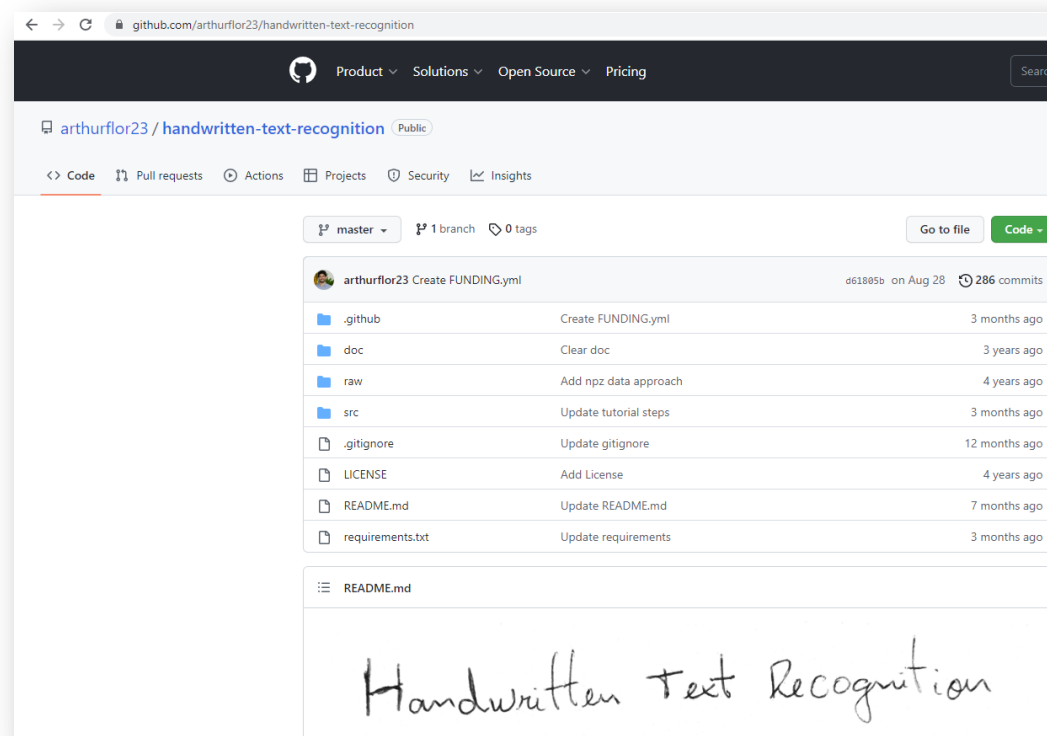


Transkribering i Transkribus

- Vi kom överens med ReadCoop om paketpris på transkribering
- Har gjort transkribering i Transkribus av hela porträttkatalogen
- Skapat mer ground truth i Transkribus
- Pga att Transkribus (ReadCoop) tar betalt för transkribering per sida är det hindrande för testning, för man har inte möjlighet att göra om när det blir fel

Transkribering i andra verktyg (1)

- För att träna transkribering behövs ground truth
- Vi har testat flera verktyg för att skapa ground truth, varav det som fungerat bäst har varit eScriptorium
- Vi har tränat och testat transkribering med verktyget HTR-Flor



Transkribering i andra verktyg (2)

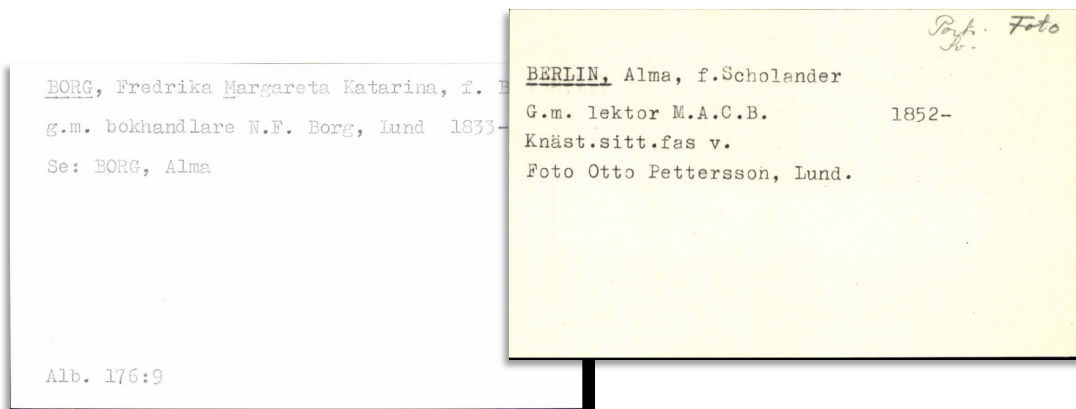
- Vi har testat segmentering och transkribering i ett verktyg som heter MMOCR
- Vi har testat träning och transkribering med Kraken / eScriptorium



MMOCR

Lagring och indexering av metadata

- Vi har indexerat den transkriberade texten i ElasticSearch (fulltextindexering)
- Den viktigaste metadatan i porträttkatalogen är "motivet"
- Vi har använt "stoppord" för att sortera bort den information som inte ingår i motivet
- Vi har förberett servern kortkatalog.ub.lu.se, men just nu finns porträttkatalogen på en annan server

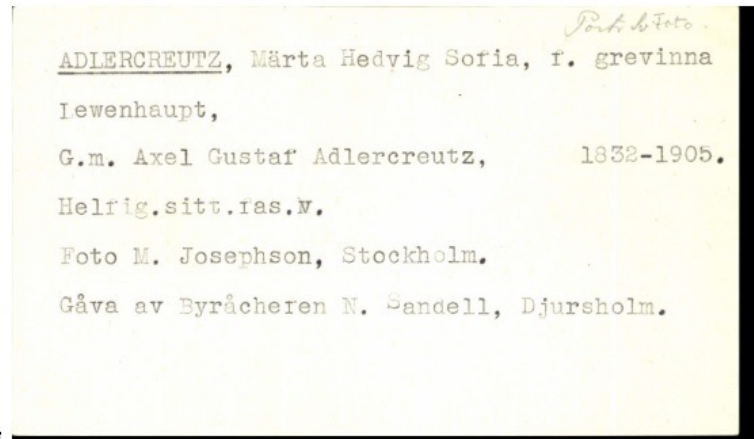


Rättning av transkriberingar

- Vi har undersökt möjligheter till batchrättningar, man kan t.ex. använda listor som
 - SBCs efternamnsindex
 - Kungalängder
 - Auktoritetslistor med Ortsnamn
- Vi kan låta kollegor göra viss rättning av transkriberingar i eScriptorium (för att göra ground truth)
- De typer av rättning vi har just nu är enbart till för kollegor, inte allmänhet

Rättning av motiv

Vi har testat motiv-rättning:
rättning om fel rader indexerats
som motiv



Portr. & Foto. ADLERCREUTZ, Märta Hedvig Sofia,
f. grevinna Lewenhaupt, 1832-1905.
G.m. Axel Gustaf Adlercreutz, Helrig.sitt.fas.N.
Foto M. Josephson, Stockholm.
Gåva av Byråcheren N. Sandell, Djursholm.



LUND
UNIVERSITY

Visning av katalog

- Vi har gjort en webbsida där man kan bläddra i och söka i katalogen
- Vi visar kortet och hela transkriberingen
- Har experimenterat med "fuzzy" sökning som gör att man kan hitta även feltranskriberade ord

PALM, Esaias
Kapten
Brb, fas h.
Foto C.V.Roikjer, Malmö.
Ur Asks saml. Album 6:63.

Sv. Album 16:63

1856-1926

PALM, Esaias
Kapten
Brb, fas h.
Foto C.V.Roikjer, Malmö.
Ur Asks saml. Album 6:63.
Portr.
Foto
Sv. Album 16:63
1856-1926

Palm, Esaias
Kapten, regementsintendent. Malmö.
1856-1924?

1 foto visitkortsformat.

Palm Esaras
Kapten regementsintendent, Malmö.
1856-1924?
foto visitkortsformat.

PALM, Esaias Laurentius
Adjunkt i Lund, fil.jubeldr
Brb, fas h.
Foto G.Gleerup, Lund.
Ur Asks album.

Portr. Foto Sv. Album 17:41

1820-1895

PALM, Esaias Laurentius
Adjunkt i Lund, fil.jubeldr
Brb, fas h.
Foto G.Gleerup, Lund.
Ur Asks album.
Foto
Portr.
Sv. Album 17:41
1820-1895



LUND
UNIVERSITY

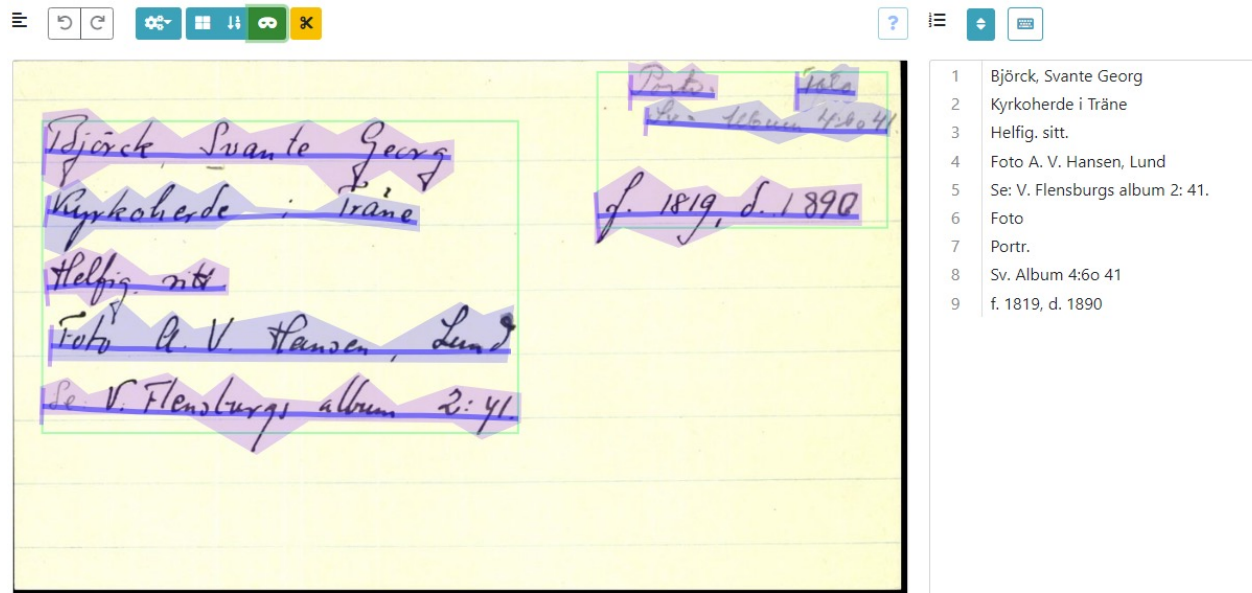
Visning av katalog, framtid

- Matchning mot ordlistor
- Ett beställningsgränssnitt ska läggas till
- Vi ska utvärdera visningen och sökfunktionaliteten
- Det ska finnas en redigeringsmöjlighet för inloggade

En viktig fråga är hur vi får användare att förstå de alternativ som finns

Våra lärdomar så långt

- En bra segmentering krävs för en bra transkribering
- Fel i ground truth för fortplantar sig i transkriberingen
- Vi kan inte träna segmentering i Transkribus
- Vi kan inte göra om transkribering i Transkribus



The screenshot displays the Transkribus web interface. On the left, a handwritten document is shown with several lines of text. The text is segmented into individual words or phrases, each highlighted with a purple background. The segments are: "Björck, Svante Georg", "Kyrkoherde i Träne", "Helfig, sitt", "Foto A. V. Hansen, Lund", and "Se V. Flensburgs album 2: 41". On the right, a list of transcription results is shown, numbered 1 through 9. The list contains the following text:

- 1 Björck, Svante Georg
- 2 Kyrkoherde i Träne
- 3 Helfig, sitt.
- 4 Foto A. V. Hansen, Lund
- 5 Se: V. Flensburgs album 2: 41.
- 6 Foto
- 7 Portr.
- 8 Sv. Album 4:60 41
- 9 f. 1819, d. 1890

Länkar

- Kat-57 idag
<http://kat57.ub.lu.se/alfabetisk/ipac/SearchForm.jsp>
- Transkribus
<https://readcoop.eu/transkribus/>
- HTR-flor
<https://github.com/arthurflor23/handwritten-text-recognition>
- MMOCR
<https://mmocr.readthedocs.io/en/latest/>
- eScriptorium
<https://gitlab.com/scripta/escriptorium>
- Riksarkivets HTR info
<https://riksarkivet.se/htr>



LUND
UNIVERSITY