ENCCS and Vega in Support of Large-Scale Digitization of Archival Materials

Riksarkivet

Digitization of Archival Data

- The Swedish National Archives has about 80000 shelf meters of archival material, ranging from the 11th century until the present
- Our digital archives consists of about 215 million scanned images of archival documents.
- Just scanning the documents don’t make them searchable. For running-text documents HTR/OCR is needed for a full-text search, and for forms and tables, segmentation and HTR/OCR is needed to transform the scanned forms into searchable databases to be used by us internally, or by research

Riksarkivet

A Few Notes on OCR/HTR

- It consists of segmentation/text-detection plus text-recognition
- In 2015 an architecture for text-recognition called CRNN (Convolutional Recurrent Neural Network) was introduced in a paper called ”An End-to-End Trainable Neural Network for Image-based SequenceRecognition and Its Application to Scene Text Recognition.” This architecture is still widely used, for example in Tesseract and ABBY for OCR, and by Transkribus for HTR.
- Since 2019 several architectures for text-recognition based on attention and the tranformer architecture, rather than recurrence, have been developed in research and implemented in open-source repositories such. The usual advantage of attention vs recurrence, that is, handling long-range context, also applies to text-recognition.
- Which architecture to choose is largely determined by the type of material you’re going to analyze

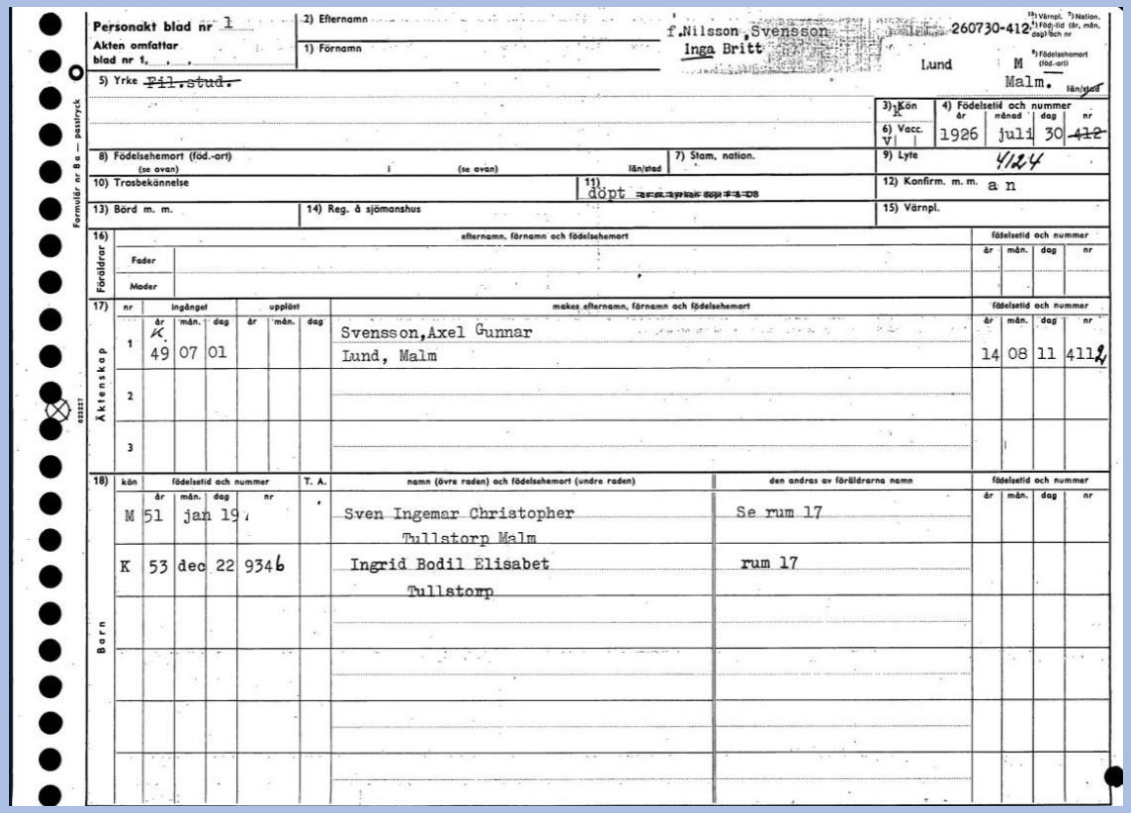
Indexing Projects at The Swedish National Archives

- The idea behind what we call indexing projects is to work with certain archival records that gets a lot of requests from the public, for example personal records or property records, and index them by some unique information, for example personal registration number or property name.
- By doing this the scanned images becomes searchable by this information, and the relevant personal record or property record can be quickly looked up.
- This streamlines our internal case handling and enhance the availability of the records

Riksarkivet

Prototype Project Started in 2021

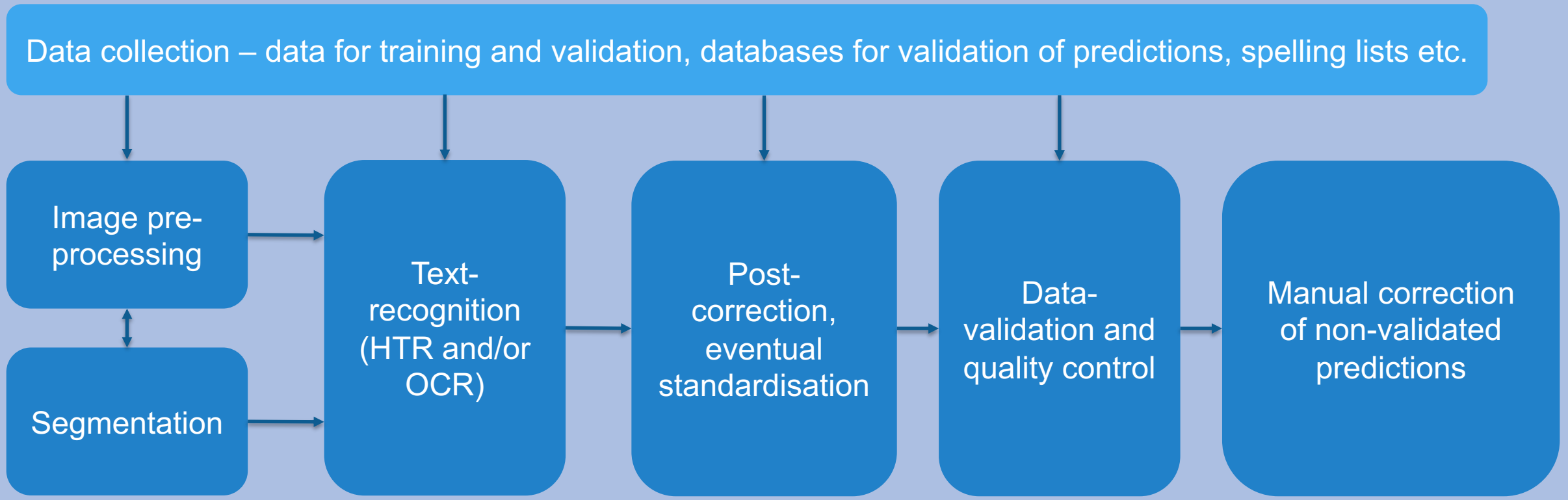
- The goal with this project was to investigate whether there is a general method and workflow for automatically indexing archival records with the help of machine-learning methods
- Our case-study was the church archive's personal records
- We didn't have a ML-infrastructure in place, and we didn't have access to any HPC-environment, so we trained models on relatively small datasets, and analyzed just a small sample of the records, with the goal of arriving at a proof-of-concept
- The results were promising and it led us to hypothesize a general method for similar indexing projects at the National Archives


Personakt blad nr 1
Akten omfattar blad nr 1, , ,
2) Efternamn f.Nilsson, Svensson
1) Förnamn Inga Britt
260730-412
Lund
M
Malm.
5) Yrke Fil.stud.
3) Kön K
4) Födelsetid och nummer år 1926 månad juli dag 30 nr 412
6) Vacc. V
8) Födelsehemort (föd.-ort) (se ovan) (se ovan) län/stad
7) Stam. nation.
9) Lyte 4/24
10) Trosbekännelse
11) döpt
12) Konfirm. m. m. a n
13) Börd m. m.
14) Reg. å sjömanshus
15) Värnpl.
16) Föräldrar — efternamn, förnamn och födelsehemort — födelsetid och nummer
Fader
Moder
17) Aktenskap — nr — ingånget år mån. dag — upplöst år mån. dag — makes efternamn, förnamn och födelsehemort — födelsetid och nummer
1 49 07 01 Svensson, Axel Gunnar, Lund, Malm 14 08 11 4112
2
3
18) Barn — kön — födelsetid och nummer — T. A. — namn (övre raden) och födelsehemort (undre raden) — den andras av föräldrarna namn — födelsetid och nummer
M 51 jan 19 Sven Ingemar Christopher, Tullstorp Malm — Se rum 17
K 53 dec 22 9346 Ingrid Bodil Elisabet, Tullstorp — rum 17

Example of a personal record (The project manager Catharina Dahlgren's grand-mother, no GDPR problem)

Riksarkivet

General indexing pipeline and workflow



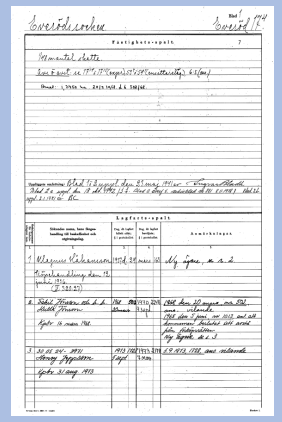
Riksarkivet

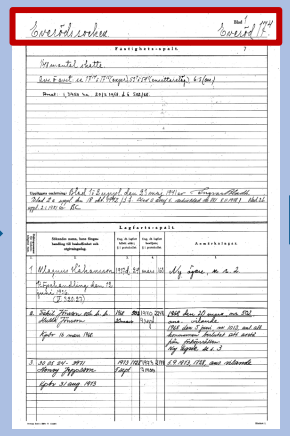
Implementing the pipeline in a production project

- Many requests from the the public made to the Swedish National Archives concerns property records. These requests requires a lot of manual work by archivists, which could be greatly reduced if the scanned images where indexed by property name
- We have approximately 18 million scanned images of property records
- About 9 million of them are already manually indexed during a period of ten years, which means we have all the training data we need.

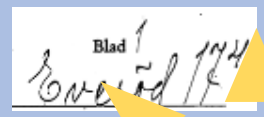
Riksarkivet

The Property Record Indexing Pipeline



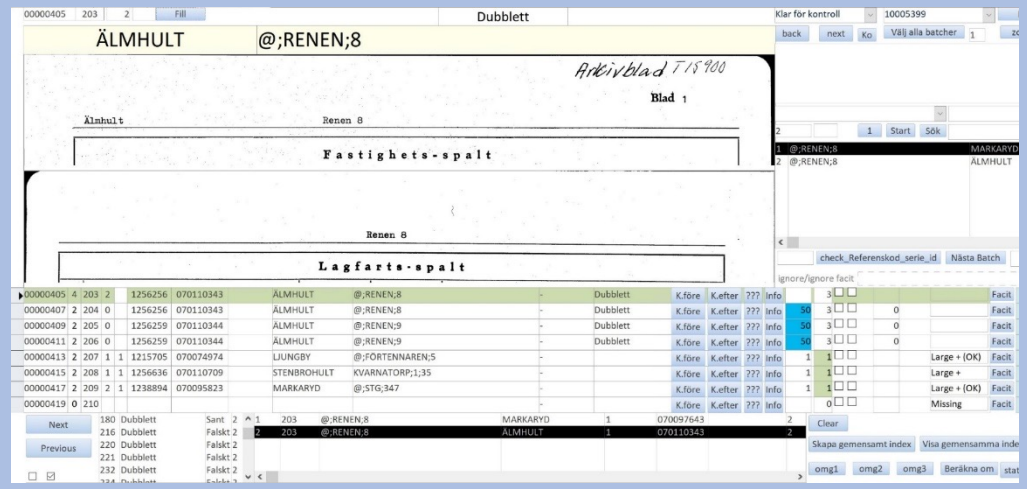


Cascade RCNN Object Detection



SATRN HTR/OCR Model

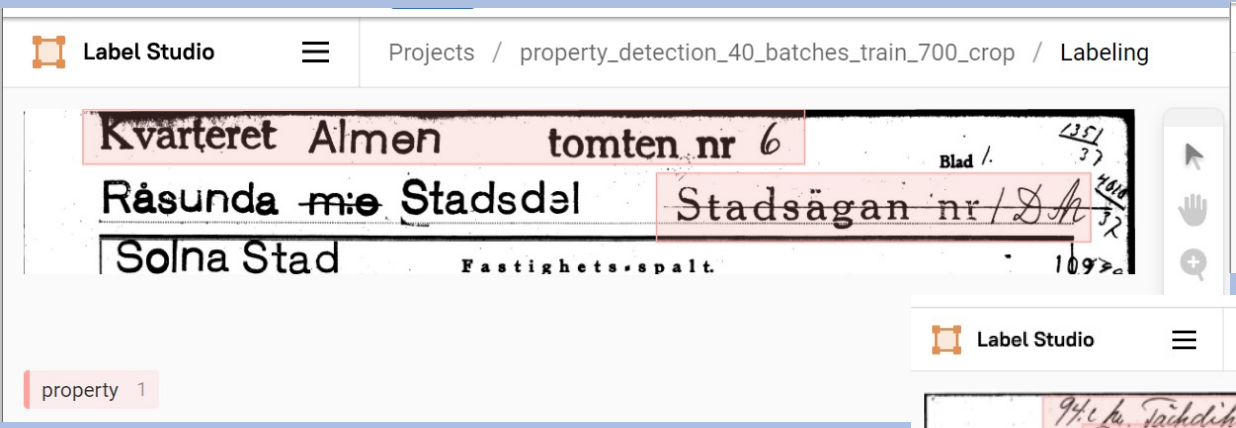
Everöd;17;4
Link against validation database

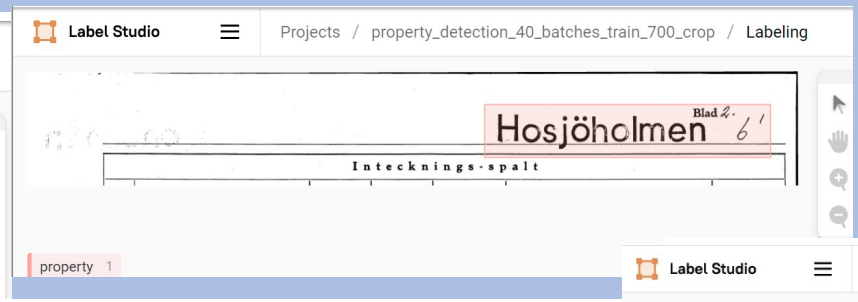


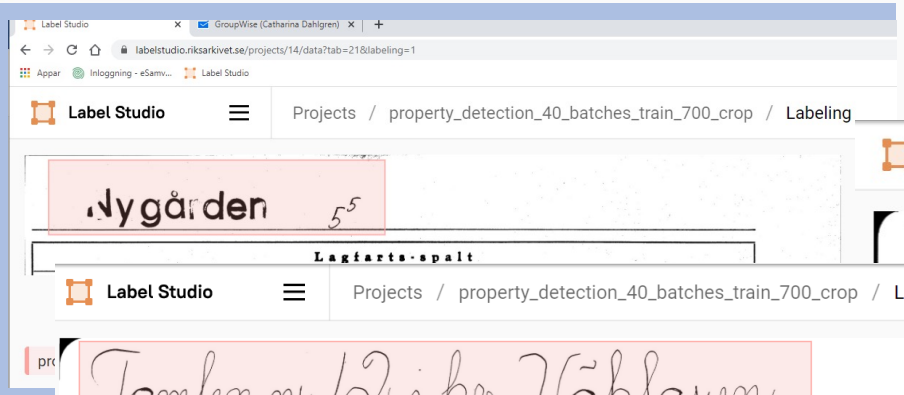
Indexing-Database

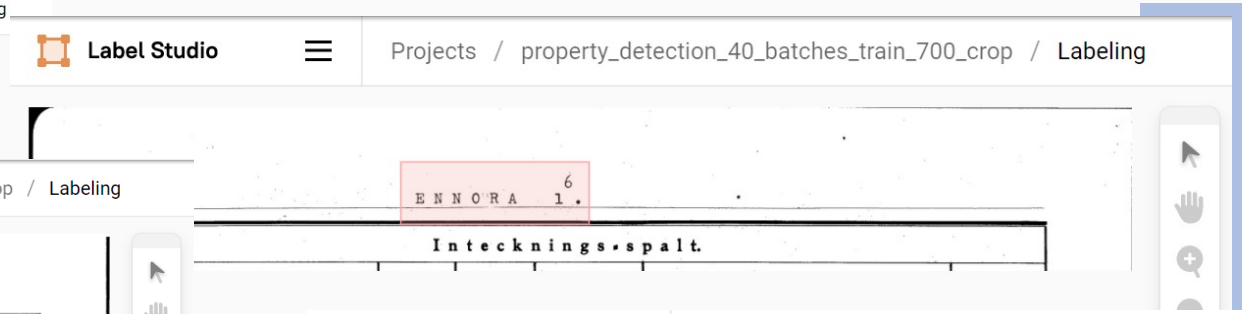
Riksarkivet

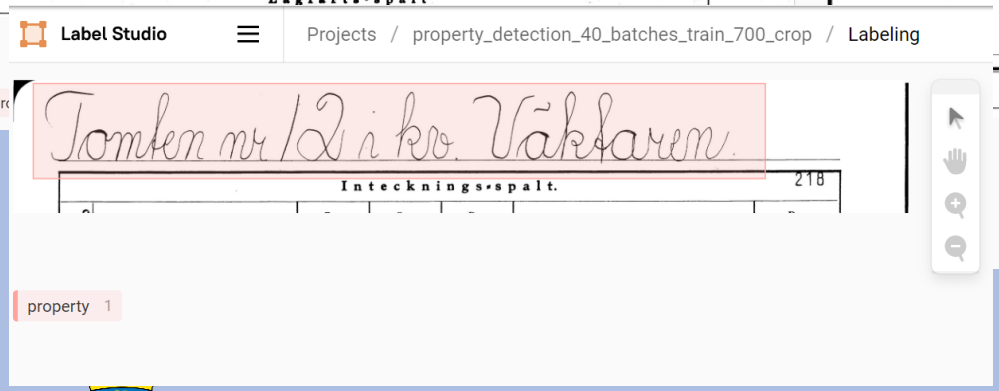
Label Studio and MMDetection for Object Detection

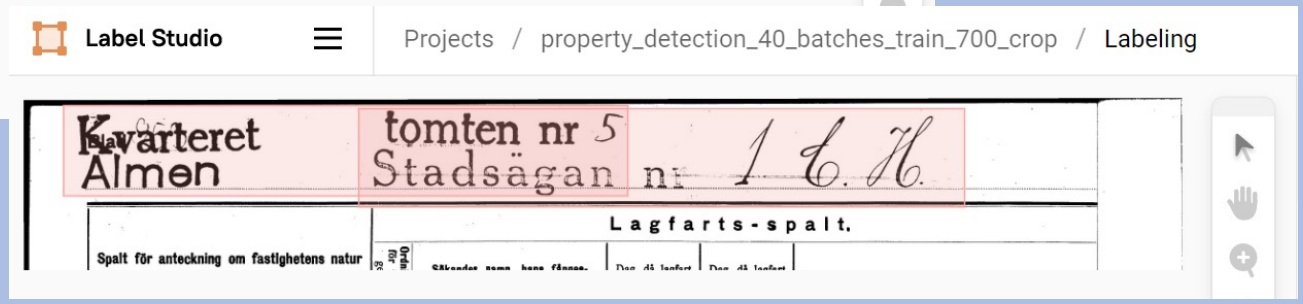












Riksarkivet

Training the Text-Recognition Model

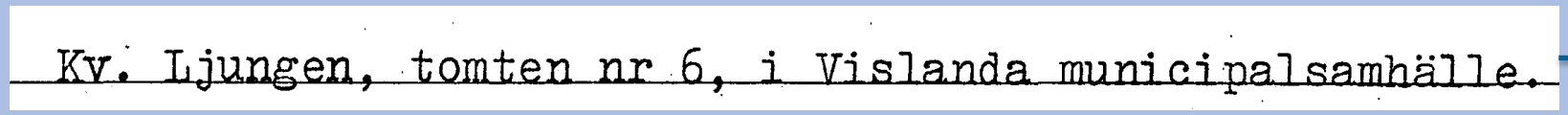
MMOCR, SATRN-Architecture

HTR/ OCR-model

Using our object-detection model, and the database for the already indexed property records we created the training set for the text-recognition model. We trained it on Vega, on approximately one million images

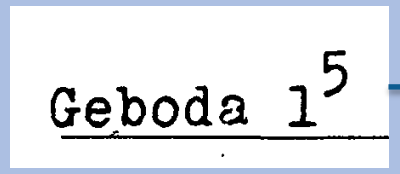
Riksarkivet

Why Attention Rather than Recurrence?

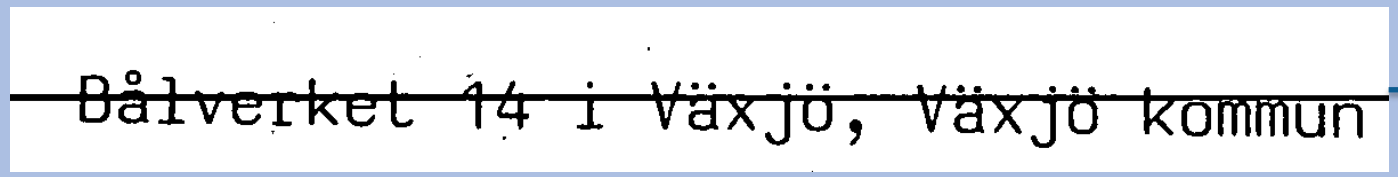


*;ljungen;6

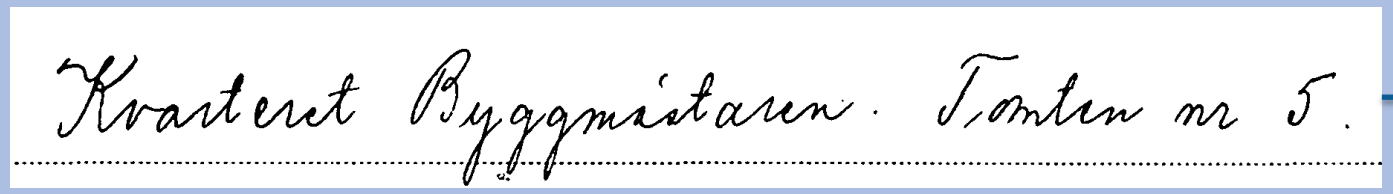
hjulatorp;8;1



geboda;1;5



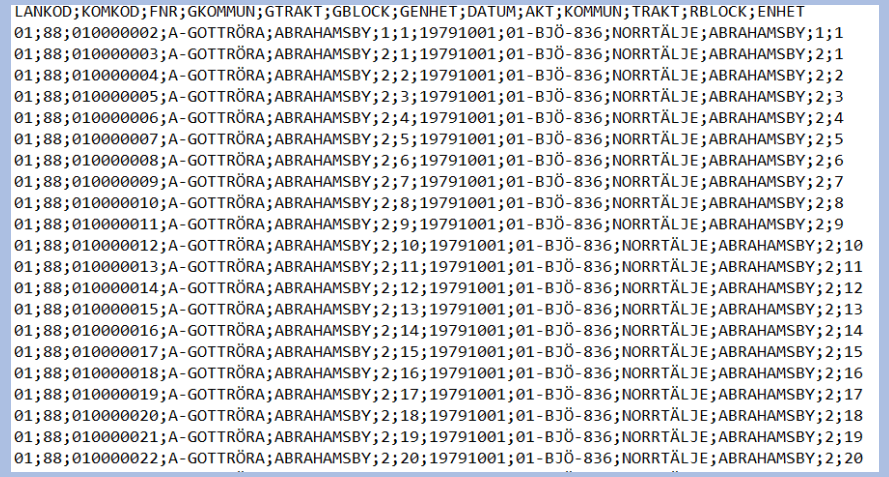
*;bålverket;14



*;byggmästaren;5

Riksarkivet

Post-Correction and Data Validation



```
LANKOD;KOMKOD;FNR;GKOMMUN;GTRAKT;GBLOCK;GENHET;DATUM;AKT;KOMMUN;TRAKT;RBLOCK;ENHET
01;88;010000002;A-GOTTRÖRA;ABRAHAMSBY;1;1;19791001;01-BJÖ-836;NORRTÄLJE;ABRAHAMSBY;1;1
01;88;010000003;A-GOTTRÖRA;ABRAHAMSBY;2;1;19791001;01-BJÖ-836;NORRTÄLJE;ABRAHAMSBY;2;1
01;88;010000004;A-GOTTRÖRA;ABRAHAMSBY;2;2;19791001;01-BJÖ-836;NORRTÄLJE;ABRAHAMSBY;2;2
01;88;010000005;A-GOTTRÖRA;ABRAHAMSBY;2;3;19791001;01-BJÖ-836;NORRTÄLJE;ABRAHAMSBY;2;3
01;88;010000006;A-GOTTRÖRA;ABRAHAMSBY;2;4;19791001;01-BJÖ-836;NORRTÄLJE;ABRAHAMSBY;2;4
01;88;010000007;A-GOTTRÖRA;ABRAHAMSBY;2;5;19791001;01-BJÖ-836;NORRTÄLJE;ABRAHAMSBY;2;5
01;88;010000008;A-GOTTRÖRA;ABRAHAMSBY;2;6;19791001;01-BJÖ-836;NORRTÄLJE;ABRAHAMSBY;2;6
01;88;010000009;A-GOTTRÖRA;ABRAHAMSBY;2;7;19791001;01-BJÖ-836;NORRTÄLJE;ABRAHAMSBY;2;7
01;88;010000010;A-GOTTRÖRA;ABRAHAMSBY;2;8;19791001;01-BJÖ-836;NORRTÄLJE;ABRAHAMSBY;2;8
01;88;010000011;A-GOTTRÖRA;ABRAHAMSBY;2;9;19791001;01-BJÖ-836;NORRTÄLJE;ABRAHAMSBY;2;9
01;88;010000012;A-GOTTRÖRA;ABRAHAMSBY;2;10;19791001;01-BJÖ-836;NORRTÄLJE;ABRAHAMSBY;2;10
01;88;010000013;A-GOTTRÖRA;ABRAHAMSBY;2;11;19791001;01-BJÖ-836;NORRTÄLJE;ABRAHAMSBY;2;11
01;88;010000014;A-GOTTRÖRA;ABRAHAMSBY;2;12;19791001;01-BJÖ-836;NORRTÄLJE;ABRAHAMSBY;2;12
01;88;010000015;A-GOTTRÖRA;ABRAHAMSBY;2;13;19791001;01-BJÖ-836;NORRTÄLJE;ABRAHAMSBY;2;13
01;88;010000016;A-GOTTRÖRA;ABRAHAMSBY;2;14;19791001;01-BJÖ-836;NORRTÄLJE;ABRAHAMSBY;2;14
01;88;010000017;A-GOTTRÖRA;ABRAHAMSBY;2;15;19791001;01-BJÖ-836;NORRTÄLJE;ABRAHAMSBY;2;15
01;88;010000018;A-GOTTRÖRA;ABRAHAMSBY;2;16;19791001;01-BJÖ-836;NORRTÄLJE;ABRAHAMSBY;2;16
01;88;010000019;A-GOTTRÖRA;ABRAHAMSBY;2;17;19791001;01-BJÖ-836;NORRTÄLJE;ABRAHAMSBY;2;17
01;88;010000020;A-GOTTRÖRA;ABRAHAMSBY;2;18;19791001;01-BJÖ-836;NORRTÄLJE;ABRAHAMSBY;2;18
01;88;010000021;A-GOTTRÖRA;ABRAHAMSBY;2;19;19791001;01-BJÖ-836;NORRTÄLJE;ABRAHAMSBY;2;19
01;88;010000022;A-GOTTRÖRA;ABRAHAMSBY;2;20;19791001;01-BJÖ-836;NORRTÄLJE;ABRAHAMSBY;2;20
```

Various other validation criteria were also applied

If validation fails, the image gets flagged for manual control in an indexing program developed for this specific purpose

We validated the predictions against an existing database of property names, to make sure the prediction is a valid property name. No guarantee against false positives.

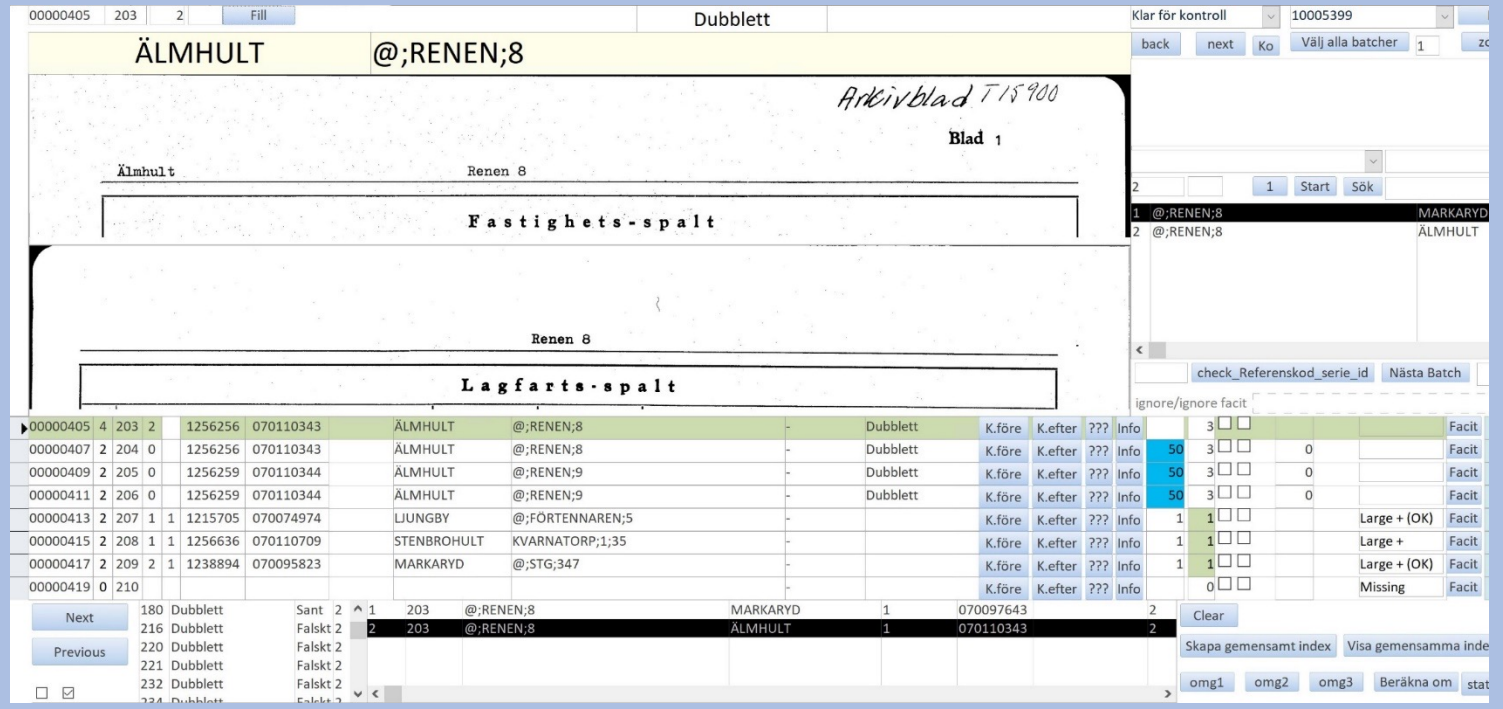
Riksarkivet

Manual Correction for Non-Validated Predictions

Manual correction where:

- Link agains validation database fails
- Low prediction-confidence
- Valid property name, but wrong region

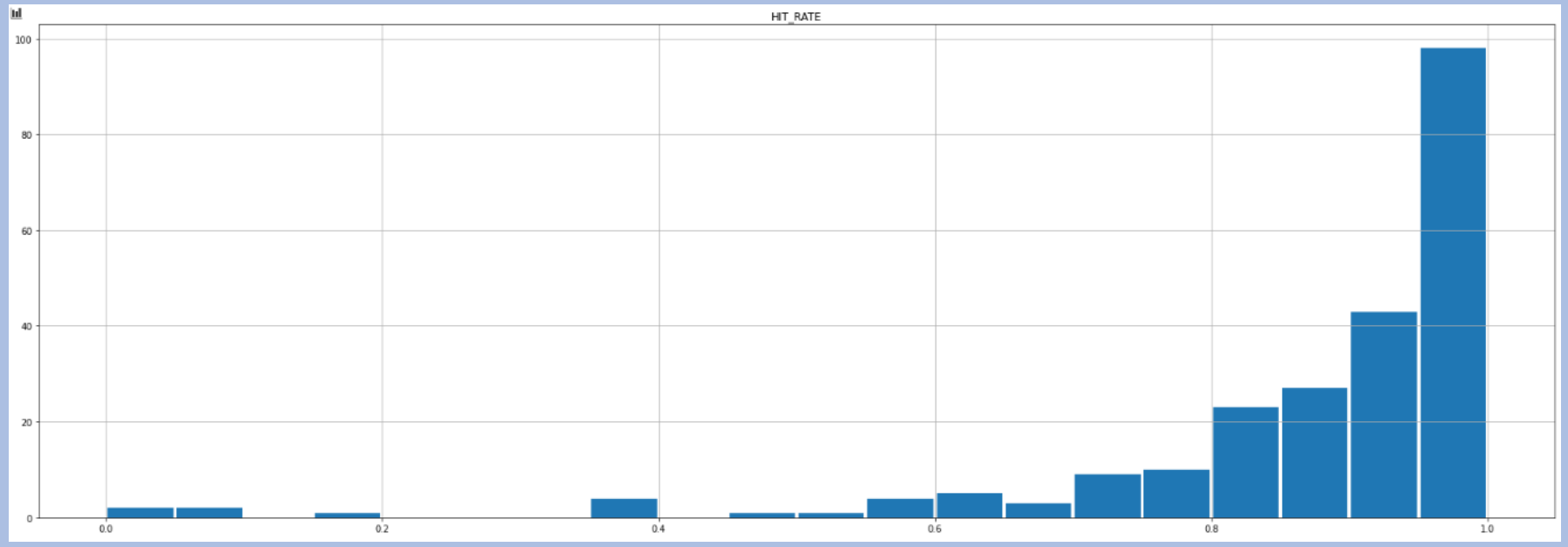
Tradeoff between number of false positives and number of hits.



Riksarkivet

Results

- We evaluated the pipeline on a testset of 100000 images spread over the data-space (locations in Sweden where the records were made). We got exact matches on just over 90% of the property names.
- Below is a venn diagram of the 233 batches the testset consisted of. Each batch consists of roughly 500 images.The bars represent 5 percent intervals. Most of the batches below 70% were actually errors in ground truth, that is, errors in the manual indexing.



Riksarkivet

Vega for training and inference

- Training the SATRN-model on Vega enabled us to increase the scale of the resized images going into the model, thereby improving accuracy for handwritten text, which generally requires more information than printed text
- Running 9 million images thorugh the pipeline on VEGA took roughly 90 node-hours
- At a hit-rate of 90% this project saves us about 700000 euros in manual labor costs, and the indexing database gets created a lot quicker

ENCCS and Vega are opening up huge possibilities

- More of the same, indexing projects that saves tax-money and streamlines our case-handling
- We are also in the process of creating running-text HTR base-models for different time-periods with the goal of mass-digitizing historical handwritten archival documents and making them available for research and for the public
- We want to improve our OCR-pipeline and apply it on scale
- We would also like to create base-models for segmentation of historical forms and tables
- One possible future project is to automatically transcribe the 19th-century censuses, this project would be a part of an already existing research infrastructure called swed-pop, of which the National Archives is a part, and where we are doing this transcriptions manually
- Inspired by the Royal Library of Sweden, we've also done experiments with training historical LM:s, that is, language models like BERT, adapted to historical text.

Thank you for listening,

And big thanks to:

ENCCS,

EuroHPC,

and Vega

Riksarkivet