What do I want to show?

We know how to extract information and structure it from the job ads in newspapers.

Steps to do it

Docling – from IBM – takes newspapers and breaks them up into components.

This tech is improving rapidly – what used to involve Image > bounding boxes > text > structure > advert we can use. Now segments images > single step extract and structure.

Steps:

1. Process images of newspapers – improve contrast and dewarp
2. Break up the adverts into their own sections
   1. Manually label the bounding boxes
   2. Draw on melissa Dell’s American stories dataset.
3. Turn images of ads into a structured markdown document with docling
4. Extract the skills.

Draw on pipeline from Arbetsformedlingen for how to get the

Talk about experience of structuring biographies – cut down the time involved with structuring biographies by X – savings both in time and money.

This is the time to do this work.

Note that Docling and off the shelf software make this task feasible in a short amount of time.

Description of NER on skills.

Use LLM to get started and cut down on labelling time with Explosion tool.

Tie to question we can answer like – “what skills were asked for over time, how did this change” – refernce to paper using job tech data on skills in private and public sector, with the lag for the public sector.