pattern recognition group

Feb 24, 2026

PSE Projekt Dodis & PRG

Dodis

- Dodis is the research center for the documentation and study of Swiss foreign policy history.
- Dodis collects and publishes archival sources from the Swiss Federal Archives concerning Swiss foreign policy, political processes, and administrative decision-making.
- ~12 employees, Civis, Hiwis; since 1997; database www.dodis.ch; release a printed edition every year; ongoing effort;
- Raw documents (Handwritten/Typewritten) from the 1848 – 1994
- Digitalization was/is a manual process, now turning into an automated process (ongoing)

Demo

- fancy-ml.dodis.ch

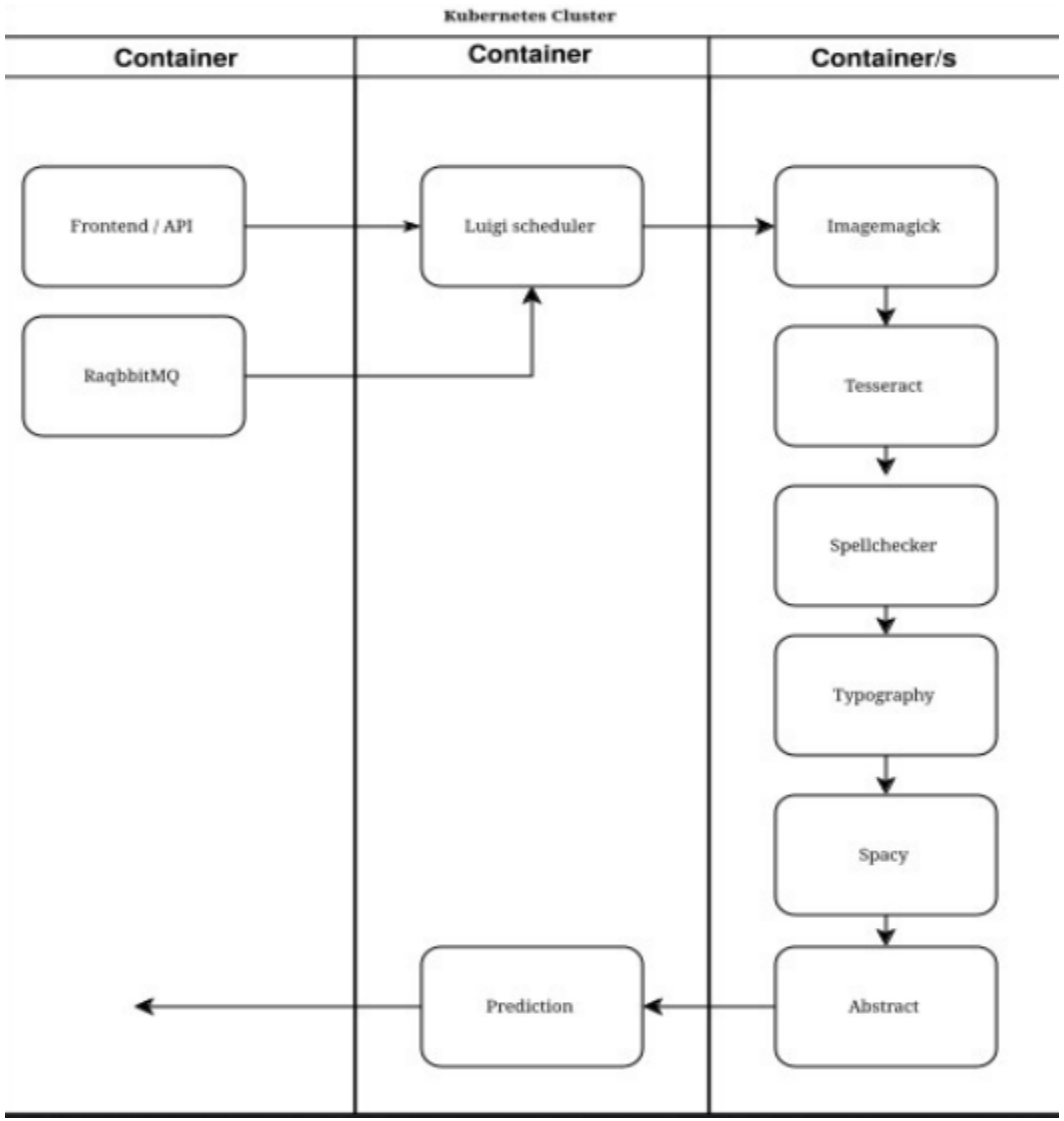
Status of today

- Working luigi pipeline with NER and NEL component
  - Running in containers on a K8
- In production since Okt. 2024
- NER/NEL component is trained on part of the Dodis TEI XML and shows mixed results
  - NER is OK
  - NEL is okayish

Model trained on Dodis data			
	P	F	R
LOC	89.51	93.11	91.27
PER	88.63	87.00	87.81
ORG	74.19	68.24	71.09

Ovall picture – fancy ml

- Pipeline with spotify/luigi to automate the transcription process
- Upload
  - Image processing
  - OCR (mistral.ai)
  - NER (spacy)
  - NEL (spacy)
  - TEI XML (python code)
  - Abstract openai



Goal

- Presentation at the KSZE conference in Bern 2026 and opening the project for more research groups
- A more flexible pipeline with several NEL component (different training)

Deliverable

- Wikidata NEL component
- Custom TEI NEL trainer
- (Improved Dodis NEL component with metadata data from the database)

In-Scope

- Improve the NEL component of the pipeline
- Deliver several or one component with different Linking targets
  - Weights
  - Code to train the pipeline
  - Readme to understand what is happening
- Write a blogpost/recipe how to train the pipeline with TEI xml's

Out Scope

- Frontend integration
- Integration into full Luigi pipeline
- Containerization of the project
- Helm chart
- Any other component of the pipeline

Tools / API

- Spacy.io as the tool to make NER/NEL
  - If you like to use an alternative you need to have good arguments
- TEI XML as training input for the custom pipeline
  - It's weird…
- Input to the component is OCR output and a language parameter
  - OCR output can be noisy
  - Documents can be multilingual (ignored in the current setup)
- Output of the component is the annotated text
  - Either as spacy doc or in another machine-readable format
- In our experience the best model are the transformer models like de_dep_news_trf

First step

- Make yourself familiarly with spacy and its architecture
- Split work in your roles and do some research
- Run a/the wikidata pipeline and import (just) historical relevant entities (loc/per/org)
  - Understand the NER and NEL process in Spacy
  - Get familiarly with the toolchain
- Train a wikidata NEL pipeline, so we can apply them to Dodis documents
- General hint: Look at the blog post of E-editiones. Something similar…
  - https://www.e-editiones.org/posts/names-sell-named-entity-recognition-in-tei-publisher/

Future

- Train a better linking for Dodis based on TEI XML and the database
  - Implement the NEL interface of spacy (based on the knowledge of step1)
  - Train a custom NER and NEL pipeline based on the dodis data
- Make the process adaptable for other historical projects (TEI Based)
  - Generalize the code, so others project can easily reuse it to train their models
  - Publish the code and a recipe how to train and use it