Tales from the road of text to knowledge

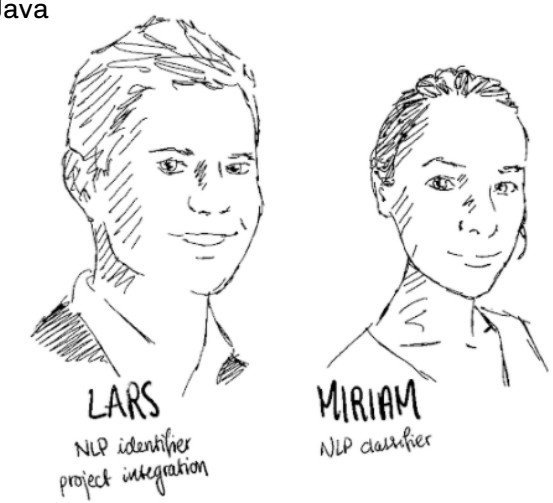
From plain text to semantic knowledge graphs
through natural language processing

By Veronika Heimsbakk

Tales from the road of text to knowledge

Veronika Heimsbakk

- Senior consultant, Capgemini Norway
- Logic & semantics, University of Oslo
- From kernel modules in C to semtech in Java



Problem

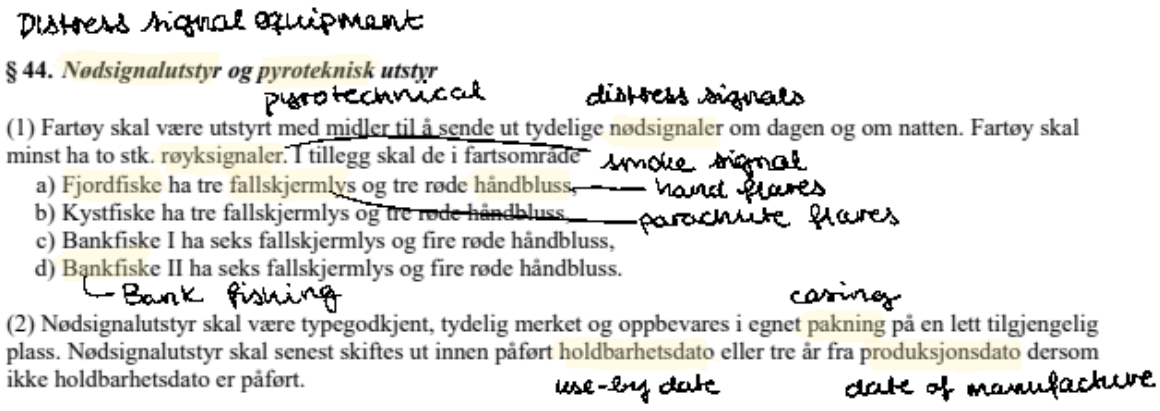
Sjøfartsdirektoratet
Norwegian Maritime Authority

- Connected and findable data in unstructured information.
- Manual modelling of graph == time consuming.

Impact

- Estimated translation time reduced by 10 000 hours.
- Minimize risk of misinterpretation.

Starting point



Distress signal equipment

§ 44. *Nødsignalutstyr og pyroteknisk utstyr*

pyrotechnical distress signals

(1) Fartøy skal være utstyrt med midler til å sende ut tydelige nødsignaler om dagen og om natten. Fartøy skal minst ha to stk. røyksignaler. I tillegg skal de i fartsområde

smoke signal

a) Fjordfiske ha tre fallskjermlys og tre røde håndbluss,

hand flares

b) Kystfiske ha tre fallskjermlys og tre røde håndbluss,

parachute flares

c) Bankfiske I ha seks fallskjermlys og fire røde håndbluss,

d) Bankfiske II ha seks fallskjermlys og fire røde håndbluss.

Bank fishing

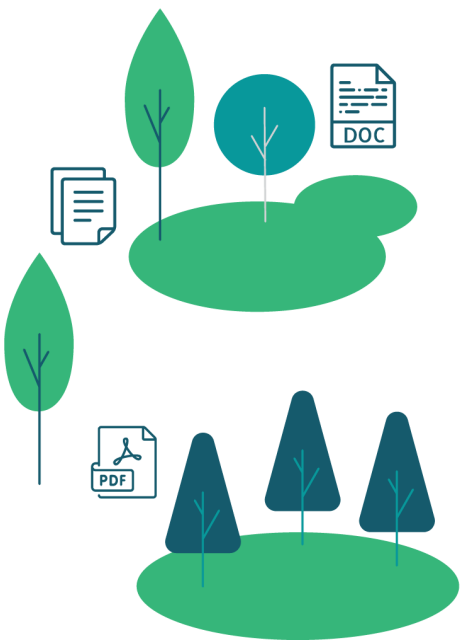
carriage

(2) Nødsignalutstyr skal være typegodkjent, tydelig merket og oppbevares i egnet pakning på en lett tilgjengelig plass. Nødsignalutstyr skal senest skiftes ut innen påført holdbarhetsdato eller tre år fra produksjonsdato dersom ikke holdbarhetsdato er påført.

use-by date date of manufacture

- Approximately 3500 triples of knowledge.
- OWL Lite ontology and SHACL shapes.

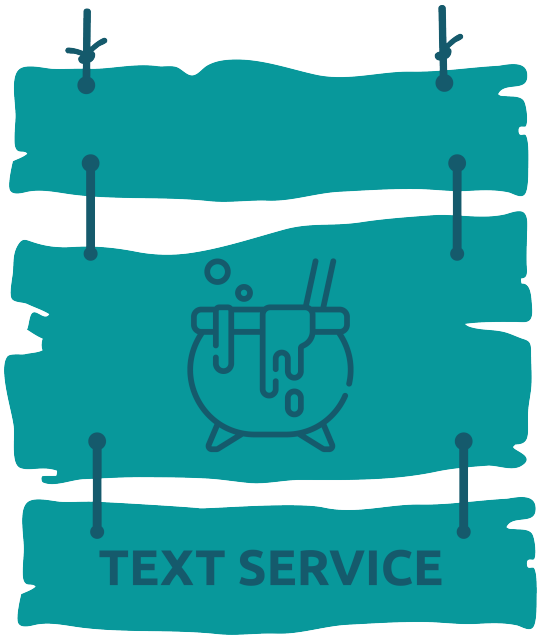
Forest of Information



- Encounter several kinds of information.
- Gather them all!

First Pitstop

- Mix information in a cauldron with a common recipe.
- Extract bits of information through API.



\regulation\chapter\paragraph\part\sub-part

Snakes & Letters

- Identifier
- Classifier

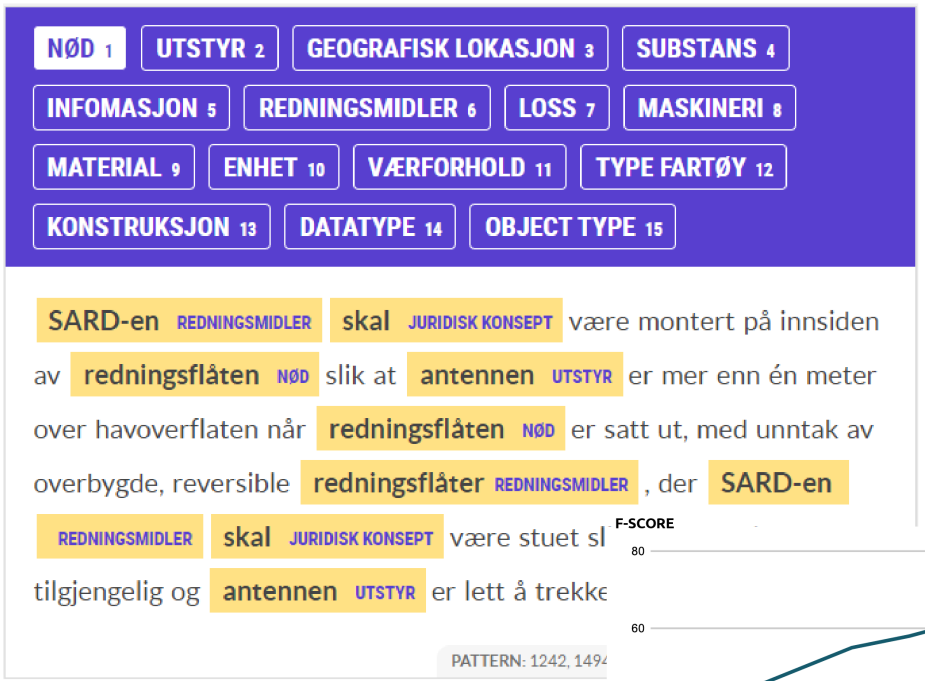
Identifier

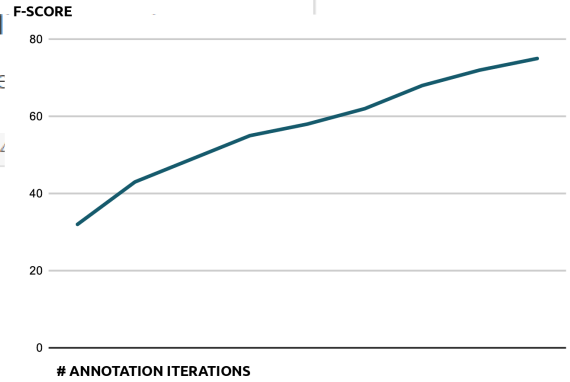
- spaCy pattern matching rules
- legal scope + location in text = regulatory requirement

```
[
  [ {"TEXT":
      {"REGEX": "^([a-zA-Z]+)(skip|fartøy|båt)$"}} ],
  [ {"LOWER": {"IN": ["lekter"]}} ],
  [ {"LOWER": {"IN": ["flyttbar"]}},
    {"LOWER": {"IN": ["innretning"]}} ]
]
```

Classifier

- Extensive annotation on a subset of information

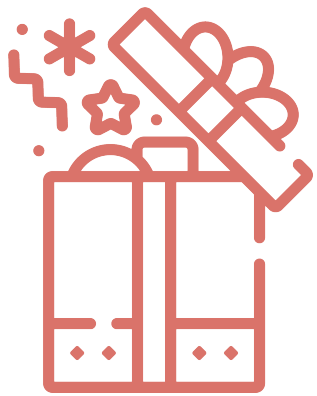




Transformation

Unboxing wonders & generate knowledge!

1. JSON-LD
2. OTTR
3. OTTR + Java
4. Java



Reasonable Ontology Templates

Template[1]

```
o-sdir:Scope[! ottr:IRI ?shape, ! ?path] :: {
    o-sh:PropertyShape(?shape, ?path),
    o-rdf:Type(?shape, sdir:Scope)
} .
```

Instance

```
o-sdir:Scope(scope:FishingVessel, sdir:vesselType) .
```

Serialized RDF

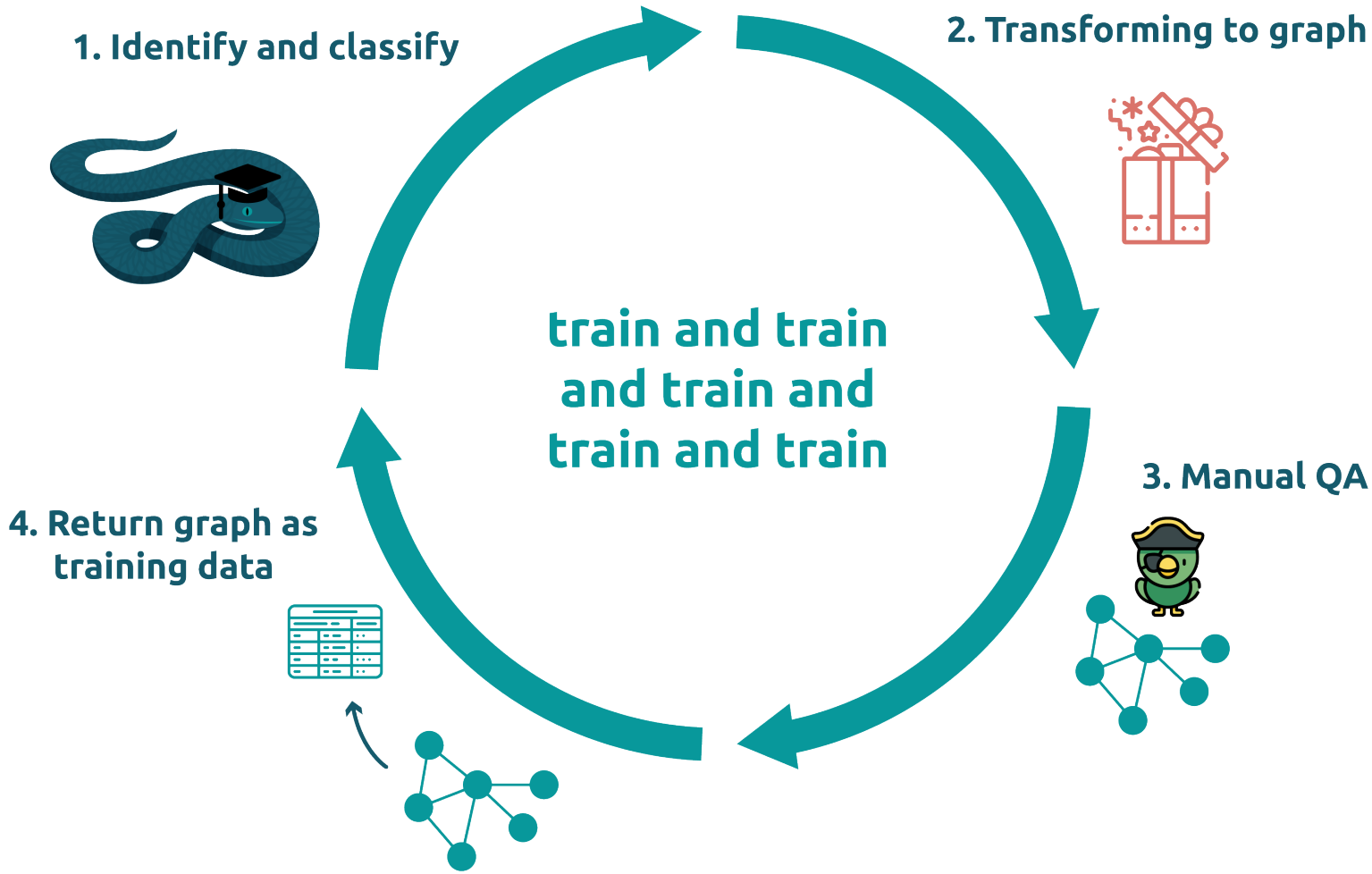
```
scope:FishingVessel a sdir:Scope, sh:PropertyShape ;
  sh:path sdir:vesselType .
```

[1] SHACL templates (`o-sh`) isn't a part of the public template library (yet).

Solution

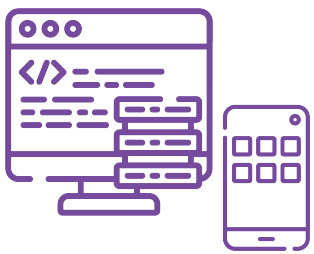
- JSON → Java object → stOTTR → RDF

Continue training



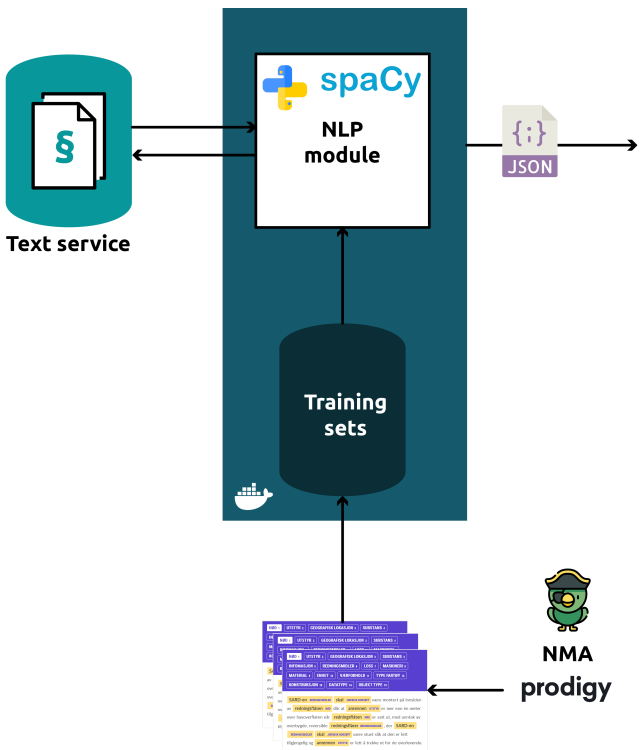
Applications

- Validation and conformance
- Maintenance of regulations
- Creating new regulations
- Search and filtering functionality

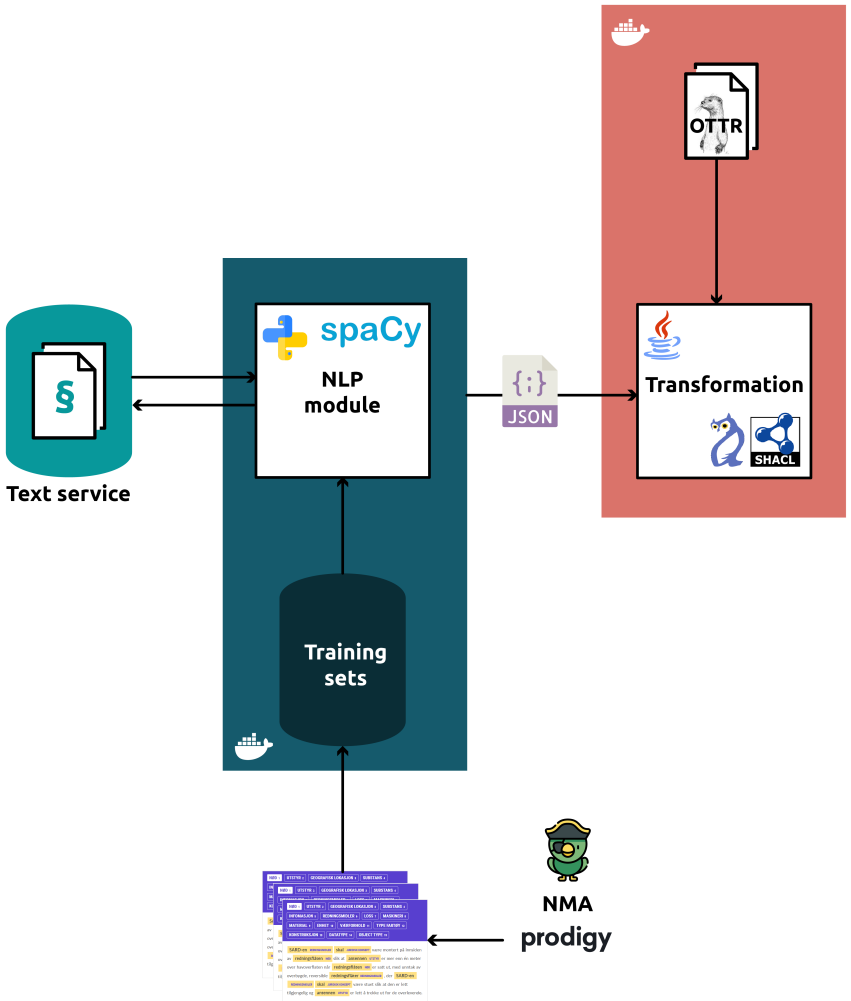


Putting the pieces together

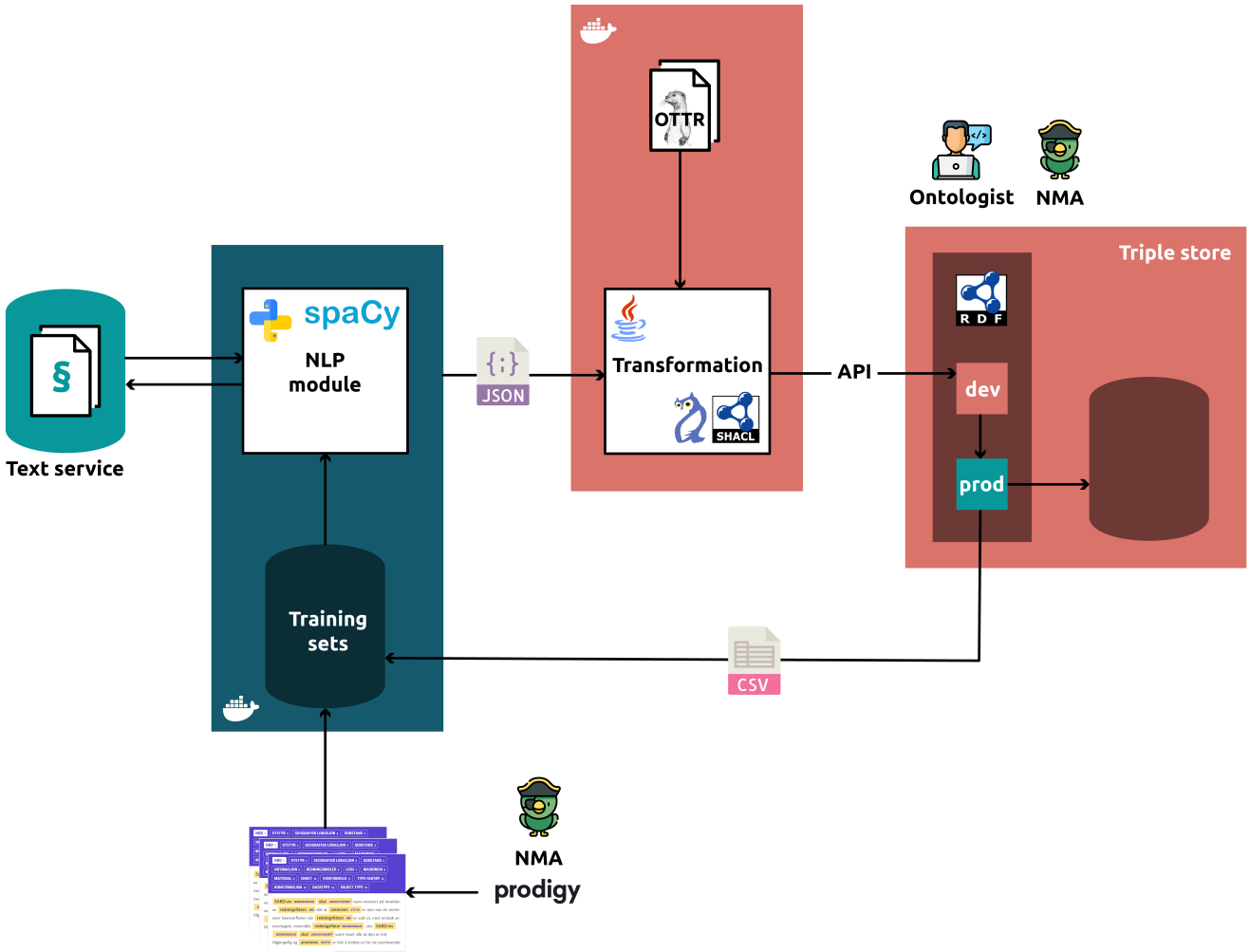
Putting the pieces together



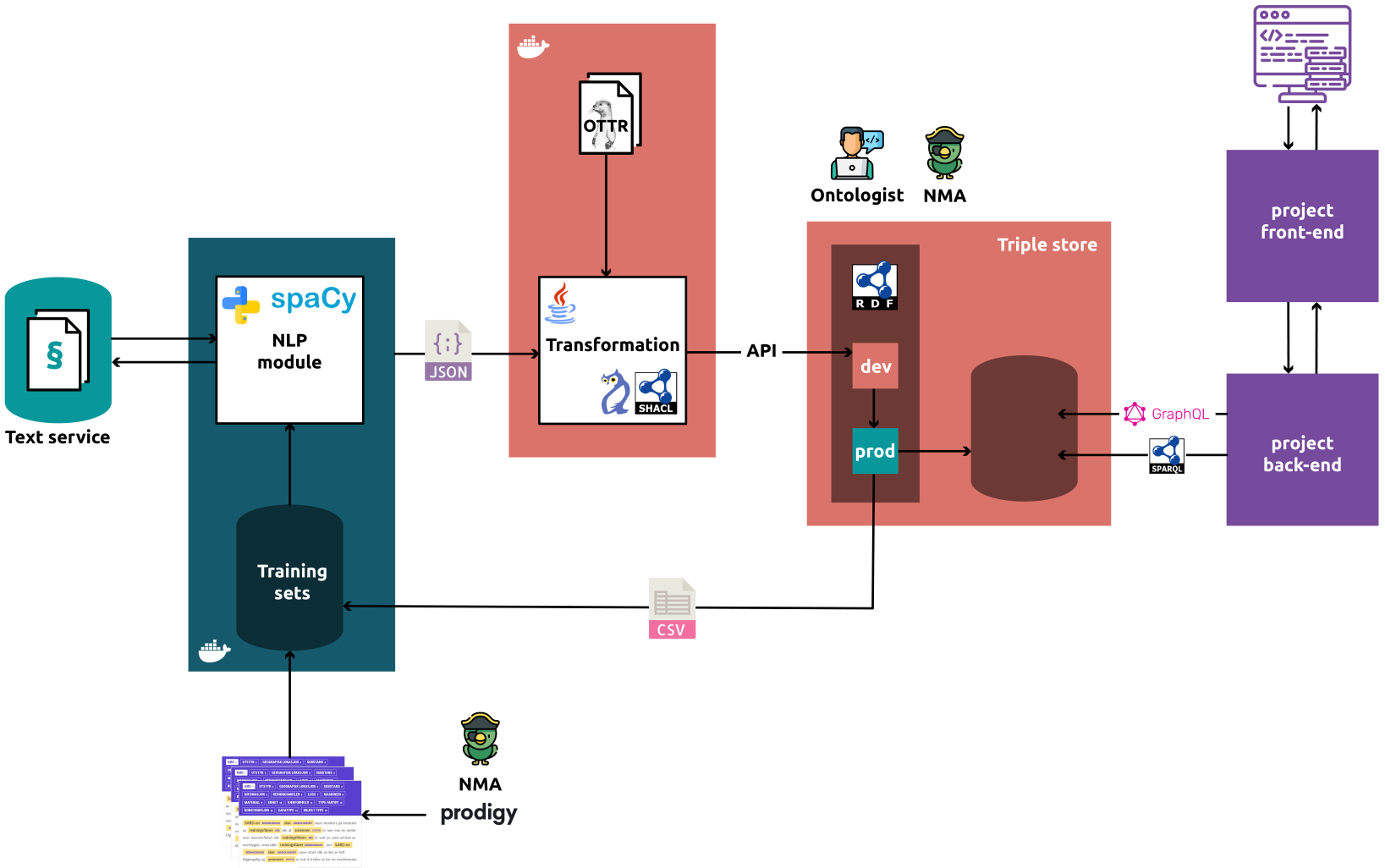
Putting the pieces together



Putting the pieces together



Putting the pieces together



References

- Illustrations: freepik.com, flaticon.com, Veronika Heimsbakk
- OTTR, https://ottr.xyz/
- SHACL, https://www.w3.org/TR/shacl/
- spaCy, https://spacy.io/
- prodigy, https://prodi.gy/

#KGC2021

Join the Conversation

@KGConference @veronikaheim

linkedin.com/company/the-knowldge-graph-conference/

youtube.com/playlist?list=PLAiy7NYe9U2Gjg-600CTV1HGypiF95d_D