Large-scale refinement of digital historical newspapers with named entity recognition
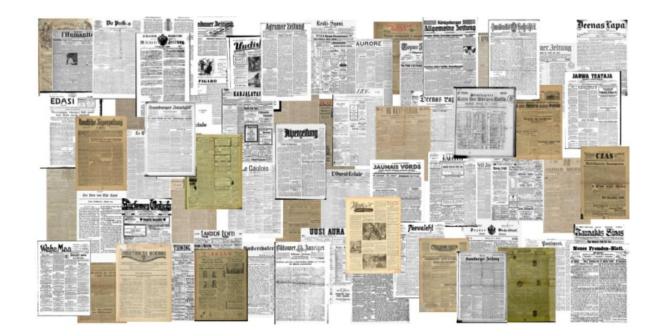
IFLA Newspaper Pre-Conference

14 August 2014, Geneva

Clemens Neudecker, SBB, *@cneudecker*

Overview

- Background
- NER Introduction
- Approach
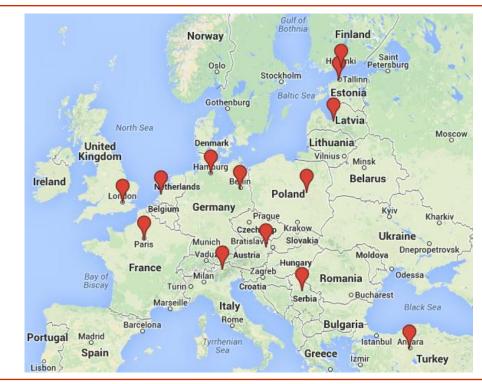- Challenges
- Scalability
- First results
- Outlook

Background

- [Europeana Newspapers](#)
  EU Best Practice Network

- **10** million newspaper pages
  with full-text from 12 libraries

- **36** million newspaper pages
  with metadata for [Europeana](#)

europeana
newspapers

ICTPSP
ICT POLICY SUPPORT PROGRAMME

Named entity recognition (I)

1. Detect names of persons, places, organisations

In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. He was visiting the United States when Adolf Hitler came to power in 1933 and did not go back to Germany, where he had been a professor at the Berlin Academy of Sciences. He settled in the U.S., becoming an American citizen in 1940. On the eve of World War II, he endorsed a letter to President Franklin D. Roosevelt alerting him to the potential development of "extremely powerful bombs of a new type" and recommending that the U.S. begin similar research. This eventually led to what would become the Manhattan Project. Einstein supported defending the Allied forces, but largely denounced using the new discovery of nuclear fission as a weapon. Later, with the British philosopher Bertrand Russell, Einstein signed the Russell–Einstein Manifesto, which highlighted the danger of nuclear weapons. Einstein was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.

Tag colours:
LOCATION TIME PERSON ORGANIZATION MONEY PERCENT DATE

europeana
newspapers

ICTPSP

Named entity recognition (II)

2. Disambiguate entities

Named entity recognition (III)

3.   Link to online resources

Approach (I)

- Tackle content in Dutch, German, French (about 50% of the 10m pages)

Approach (II)

- Use a machine learning tool (open source) developed by Stanford University, adapted for Europeana Newspapers by KBNL

  https://github.com/KBNLresearch/europeananp-ner

Approach (III)

- Create (and release) training material by manually annotating named entities on OCR'd newspaper pages

Challenges

- OCR quality

- Multiple (mixed) languages

- Historical spelling

Scalability

- Stanford NER software is multi-threaded e.g. 4 CPU cores – 4x throughput

- Optimise the NER classifier by filtering noise and sentences without NE's marked

- Robust proven Java technology

First results (Dutch)

	Persons	Locations	Organizations
Precision	0.940	0.950	0.942
Recall	0.588	0.760	0.559
F-measure	0.689	0.838	0.671

First results (French)

🇫🇷	Persons	Locations
Precision	0.529	0.548
Recall	0.834	0.216
F-measure	0.622	0.310

*   Score for organisations omitted since not enough present in the source material

europeana
newspapers

ICTPSP
ICT POLICY SUPPORT PROGRAMME

Outlook

- Q3: Release of training data for Named Entity Recognition in Dutch, German, French

- Q3: First results for German (Austrian, Italian/South Tirol), final results for Dutch, French

- Q4: Release of software (open source) for disambiguating and linking of NER results to DBPedia

www.europeana-newspapers.eu/

www.theeuropeanlibrary.org/tel4/newspapers

https://github.com/KBNLresearch/europeananp-ner

Thank you for your attention!

IFLA Newspaper Pre-Conference

14 August 2014, Geneva

Clemens Neudecker, SBB, @cneudecker