europeana
newspapers

ICTPSP
ICT POLICY SUPPORT PROGRAMME

Development of Named Entities Recognition for French Newspapers

Journée d'information
Europeana Newspapers

27/11/2014
BnF / Paris, France

Clemens Neudecker, State Library Berlin

@cneudecker

Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

What is „Named Entity Recognition“?

- Named Entity Recognition (NER) is a sub-task of Information Extraction and is typically understood as being part of the area of Computational Linguistics / Natural Language Processing.

- The main aim of NER is the automatic extraction and classification of knowledge or information from semantically unstructured text.

- NER is still subject to academic research (cf. Google & MSR Competition) – practical use in the cultural heritage digitisation sector remains a rare case.

europeana newspapers

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

ICT PSP ICT POLICY SUPPORT PROGRAMME

Asked differently: What is a „Named Entity“?

- PERSON:
  - Names of persons and families, but also names of fictional persons („Albert Einstein“, „Präsident der USA“, „Micky Maus“)
- ORGANISATION:
  - Names of companies, governemental or non-governemental organisations („IBM“, „The Beatles“, „Labour Party“)
- PLACE:
  - Cities, Provinces, Counties, geographical areas, asf. („Paris“, „Haute-Pyrénées“, „Alpes“)

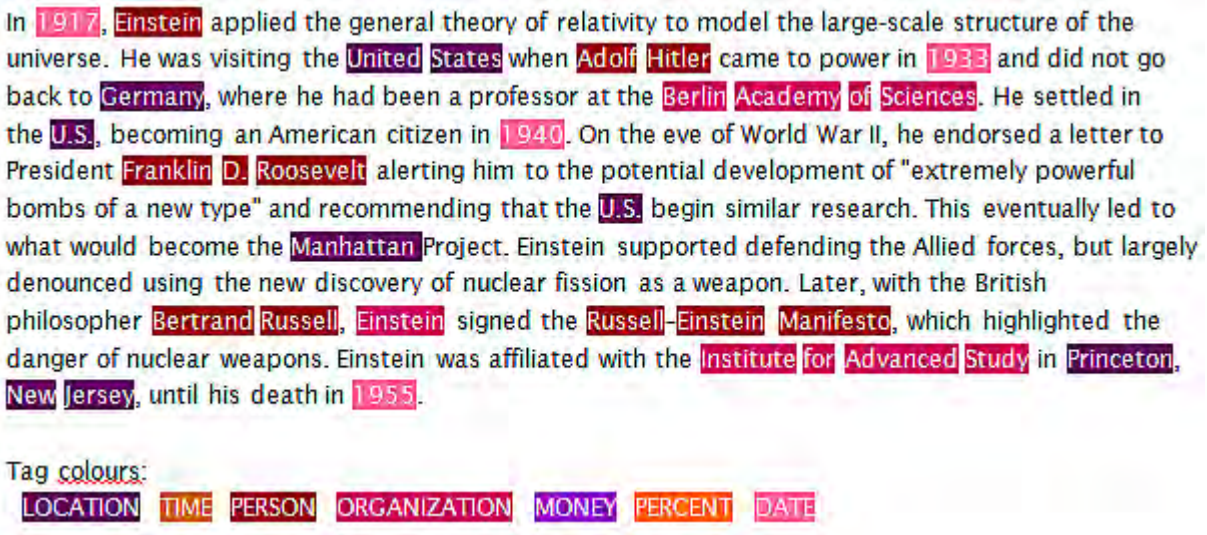
europeana newspapers

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

ICT PSP ICT POLICY SUPPORT PROGRAMME

NER (I)

1. Detection/Classification of person names, places and organisations in a running text (includes POS)



This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

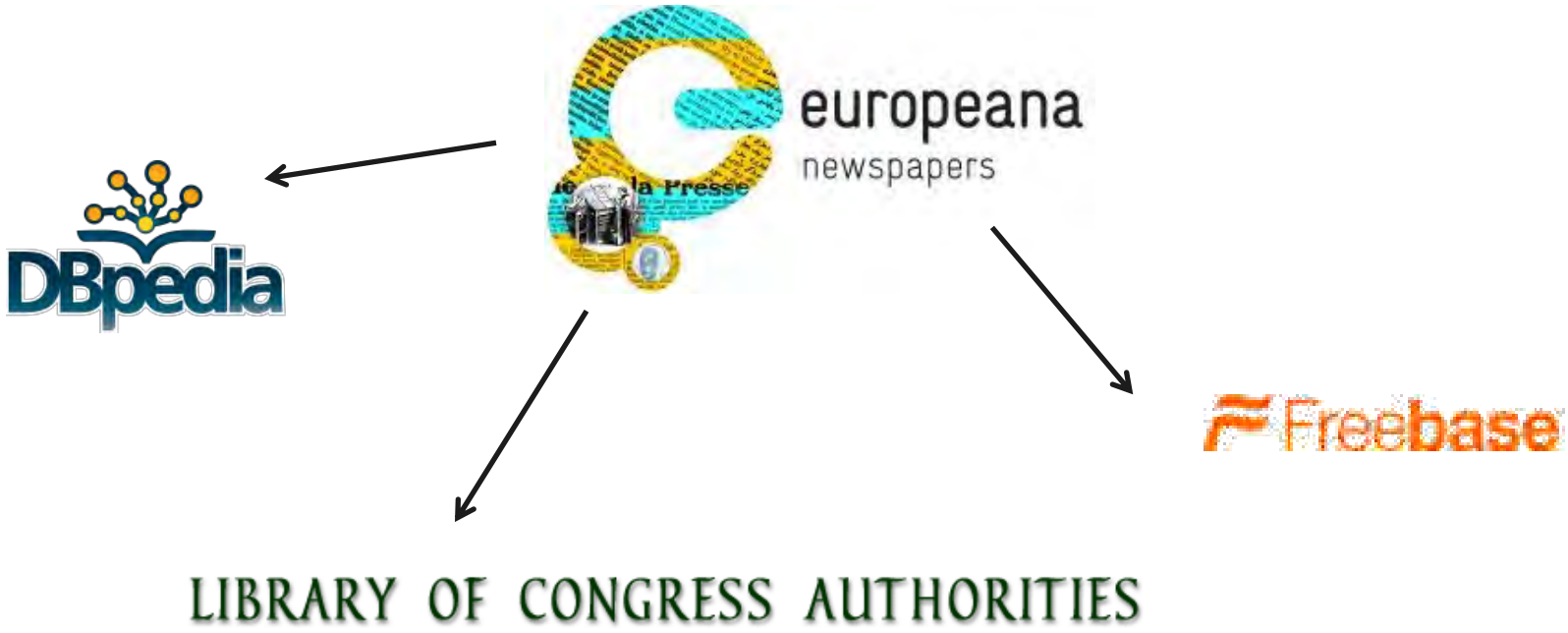
NER (II)

2. Disambiguation of terms (Example “Jordan”) through contextual information

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

NER (III)

3. Linking to authority files and online databases (Linked Data)



europeana newspapers

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

ICT PSP
ICT POLICY SUPPORT PROGRAMME

Supported languages in ENP

3 Languages:

- German
- Dutch
- French

europeana newspapers

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

ICT PSP ICT POLICY SUPPORT PROGRAMME

Approaches

- Machine learning vs. rule-based
- Advantages of machine-learning systems:
  - No need for specific linguistic expertise
  - Processing of large amounts of material
- Advantages of rule-based systems:
  - Can be tuned to very high accuracy for particular texts
  - Adaptation to local grammar and specific text style

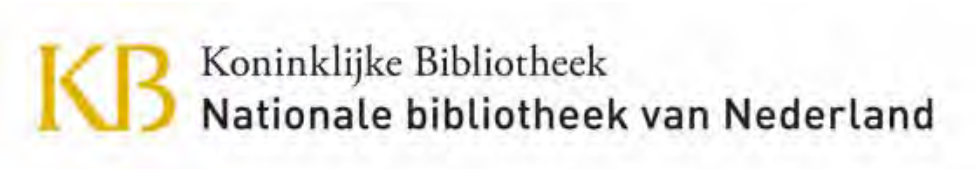
europeana newspapers

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

ICT PSP ICT POLICY SUPPORT PROGRAMME

Software

- Open Source ML software developed by Stanford University, adapted and extended for Europeana Newspapers by the KB National Library of the Netherlands



- Software is available as open source from Github for download and testing: https://github.com/KBNLresearch/europeananp-ner

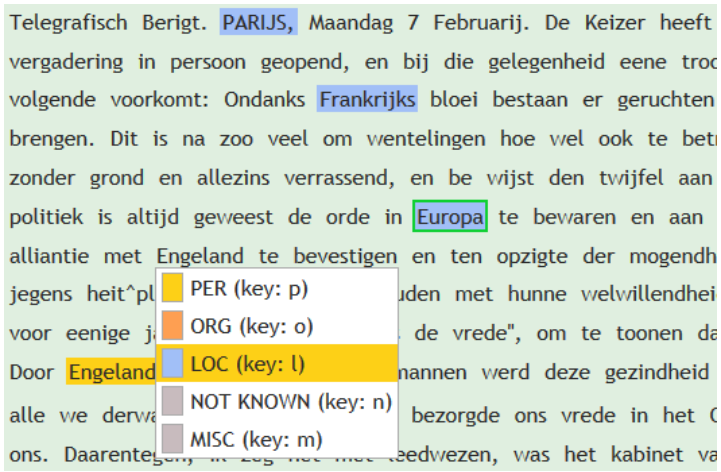
europeana newspapers

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

ICT PSP
ICT POLICY SUPPORT PROGRAMME

Training

- Training the NER systems with the help of manually annotated corpora („gold corpus“) and gazzetteers



- Publication of annotated data from ENP as open data

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Encoding

- Results of NER are stored in a library specific format: ALTO (Analyzed Layout and Text Object)

- Versions > 2.1 of ALTO specifically allow to use NER „Tags“

```
<String STYLEREFS="ID7" HEIGHT="132.0" WIDTH="570.0" HPOS="5937.0"
VPOS="3279.0" CONTENT="Reynolds" WC="0.95238096" TAGREFS="Tag5"></String>
<String STYLEREFS="ID7" HEIGHT="102.0" WIDTH="540.0" HPOS="18438.0"
VPOS="22008.0" CONTENT="Baltimore" WC="0.82539684„ TAGREFS="Tag10"></String>
…
<Tags>
    <NamedEntityTag ID="Tag5" TYPE="Person" LABEL="Reynolds"/>
    <NamedEntityTag ID="Tag6" TYPE=”Place" LABEL=”Baltimore"/>
</Tags>
```

europeana newspapers

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

ICT PSP ICT POLICY SUPPORT PROGRAMME

Problems and challenges

- OCR errors reduce the accuracy of the classification and slow down the overall processing time for recognition due to high noise.

- Historical spelling variation for person names and place names in particular.

- In many cases the historical spelling variants can not be found in online knowledge bases.

→ Specific adaptation of the software via external modules

europeana newspapers

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

ICT PSP ICT POLICY SUPPORT PROGRAMME

Initial results: Dutch

	Persons	Places	Organisations
Precision	0.940	0.950	0.942
Recall	0.588	0.760	0.559
F-measure	0.689	0.838	0.671

europeana newspapers

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

ICTPSP ICT POLICY SUPPORT PROGRAMME

Why Named Entity Recognition?

- Example: Analysis of log files from the newspaper collection of the National Library of Wales shows that *9 out of 10 queries are for a person or place name!*

(Source: Paul Gooding, Exploring Usage of Digital Newspaper Archives through Web Log Analysis: A Case Study of Welsh Newspapers Online, presented at DH2014, Lausanne)

europeana newspapers

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

ICT PSP ICT POLICY SUPPORT PROGRAMME

europeana
newspapers

ICTPSP
ICT POLICY SUPPORT PROGRAMME

Thank you for your attention!
Merci de votre attention!

@eurnews

http://www.europeana-newspapers.eu

http://www.theeuropeanlibrary.org/tel4/newspapers

http://www.europeana.eu/

Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz