# Research and development efforts on the digitized historical newspaper and journal collection of The National Library of Finland

**Kimmo Kettunen, Mika Koistinen and Teemu Ruokolainen**

**The National Library of Finland, Mikkeli unit, DH Projects**

Programme for Sustainable Growth and Jobs

Leverage from the EU 2014–2020

European Union
European Regional Development Fund

## Long history in collection building

- The National Library of Finland (NLF) has digitized historical newspapers, journals and ephemera published in Finland since the late 1990s. The present collection consists of about 12.9 million pages mainly in Finnish and Swedish.
- Out of these about 7.36 million pages are freely available on the web site digi.kansalliskirjasto.fi (Digi).
- The time period of the open collection is from 1771 to 1929.



## Research and development

- Besides producing the digitized publication data all the time NLF has been involved in research and improvement of the digitized material during the last years.
- We ended a two year European Regional Development Fund (ERDF) project in July 2017 and started another two year ERDF project in August 2017.
- Digitalia research center together with XAMK.

Programme for Sustainable Growth and Jobs
Leverage from the EU 2014–2020
European Union
European Regional Development Fund

## Research and development

- NLF is also involved in research consortium **COMHIS**, *Computational History and the Transformation of Public Discourse in Finland, 1640-1910*, which is funded by the Academy of Finland (2016–2019). COMHIS utilizes NLF's historical newspaper and journal data as one of its main sources in its research of changes of publicity in Finland.
- EU Horizon project **NewsEye** starts in May 2018. NLF is one of the partners in the project.

Programme for Sustainable Growth and Jobs
Leverage from the EU 2014–2020
European Union
European Regional Development Fund

## Data improvement and new ways to use the data: accomplishments

- Word level quality analysis for the Finnish part of data
- Open data delivery package of 1771-1910 newspapers and journals (available from digi.kansalliskirjasto.fi/opendata)
- Several improvements for the Web interface (time-line, notebook property etc.)
- Ground truth data of Finnish for new optical character recognition (open data)
- A new OCR process with Tesseract 3.04.01 (especially Fraktur font)
- Named Entity Recognition evaluation collection and trained NER model

Programme for Sustainable Growth and Jobs
Leverage from the EU 2014–2020
European Union
European Regional Development Fund

## NER – Named Entity Recognition

What?

- Names of persons, locations, organisations are important factual data in texts
- They can be recognized automatically to a reasonable extent (70-90+ %)
- They can be used in information extraction out of the data
- Names of persons and locations are used heavily as keywords in text searches of on-line databases. Many times even 80 % of keywords are names.



## An example of names in a text

**Saimaa, 22.11.1895**

Edellä sanotun tarpeen maatimana ja kun ei täällä ole mallikelpoista maanmiljelystaloa, jonka puuttce» poistamiseksi toiwomme että usei»mai»ittu koulu perustettaisiin hra O. L, Vumeruksen omistmnalle Kupiala» tilalle Rantasalmen pitäjään, eikä lärmikylään Joroisissa, koska lärmikylä jo kaumcmman aikaa nykyisen omistajansa jalomielisestä ja paljon uhraaivaiscsta toimesta un ollut, ja luonnollisesti tulisi edelleen olemaan PohjoiS-samolaisille monien kokeittcnsa ja mallikelpoisen maanmiljcll,ksen

6 names, 2 of them spelled right



## A training & evaluation set for NER

- Our complete data set consists of 170 manually annotated pages (248 544 word tokens) and 101 semi-manually annotated pages (211 034 word tokens).
- In total there are 10 457 entities of person, and 13 266 entities of location marked in our data.
- The evaluation set consists of 34 manually annotated pages. The resulting training and evaluation sections contain thus 237 and 34 pages (381 356 and 67 223 word tokens),



## NER results for historical newspaper Finnish: precision, recall, F score

TABLE I. PRECISION, RECALL, AND F-SCORES FOR EACH NAMED ENTITY CLASS ON THE GROUND TRUTH EVALUATION SET

| Class | Precision | Recall | F1 | # found | #gold standard |
|---|---|---|---|---|---|
| LOC | 0.8872 | 0.8566 | 0.8716 | 1764 | 1826 |
| PER | 0.8408 | 0.7801 | 0.8093 | 1118 | 1205 |

Ideal results with the GT data (hand corrected)

Realistic results: new Tesseract OCR decrease of 9-10% units

TABLE II. PRECISION, RECALL, AND F-SCORES FOR EACH NAMED ENTITY CLASS ON THE OCR EVALUATION SET

| Class | Precision | Recall | F1 | # found | #gold standard |
|---|---|---|---|---|---|
| LOC | 0.8527 | 0.7322 | 0.7879 | 1485 | 1826 |
| PER | 0.7856 | 0.6631 | 0.7192 | 1017 | 1205 |

## Results of a LSTM-CRF (Lample et al. 2016)

- LSTM (long short-term memory), recurrent neural network model: state-of-the-art, receives very similar results with Stanford

| Class | Precision | Recall | F1 | # found | # gold standard |
|---|---|---|---|---|---|
| LOC | 0.8598 | 0.6884 | 0.7646 | 1471 | 1826 |
| PER | 0.8212 | 0.6822 | 0.7452 | 1022 | 1205 |



## Next step

- We believe that the performance of Stanford NER with our data has reached its realistic peak (only massive amount of new training data could improve the results)
- We have more or less settled morphological disambiguation of the names after Stanford (using Omorfi+FinnPos)
- We need to figure out, how to use names in digi.kansalliskirjasto.fi → start with browsing assistancy with Uusi Suometar

Programme for Sustainable Growth and Jobs
Leverage from the EU 2014–2020
European Union
European Regional Development Fund

## Future – work in progress

- Re-OCR for the whole collection (starting with Finnish material)
- Named Entity Recognition for the Finnish material (starting with one newspaper)
- Article extraction trials from pages of one newspaper (86 000 pages)
- Simple image classification, if feasible
- GT OCR data for Swedish

Programme for Sustainable Growth and Jobs
Leverage from the EU 2014–2020
European Union
European Regional Development Fund

# Thank you for your patience!

**Kimmo Kettunen, Mika Koistinen and Teemu Ruokolainen**
**The National Library of Finland, Mikkeli unit, DH Projects**

Programme for Sustainable Growth and Jobs
Leverage from the EU 2014–2020
European Union
European Regional Development Fund