Data for digital humanities – digitized historical newspaper and journal collection of The National Library of Finland

Kimmo Kettunen
Mika Koistinen
Teemu Ruokolainen
The National Library of Finland, DH Projects

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

Digi.kansalliskirjasto.fi

Digital newspaper and journal collections of the NLF: 12 M pages 1771-

NEWSPAPERS

Digitized 5,421,841 pages.
Free use 2,954,708 pages (54%) (1771-1920).

Copyright based material 2,467,133 pages (46%) (1921-).

JOURNALS

Digitized 6,392,306 pages.
Free use 2,163,786 pages (33%) (1771-1920).

Copyright based material 4,228,520 pages (67%) (1921-).

Growing data

- The digitized collection of NLF is part of globally expanding network of library produced historical data that offers researchers and lay persons insight into past.
- In 2012 it was estimated that there were about 129 million pages and 24 000 titles of digitized newspapers in Europe (Dunning, 2012). A very conservative estimation about worldwide number of titles is 45 000 (The State of the Art, 2015).
- The number of currently available titles is probably much bigger, as the national libraries have been working steadily with digitization both in Europe, Northern America and rest of the world.

NLF's DH efforts with the newspaper and journal collection

- Besides producing and publishing the digitized raw data all the time NLF has been involved in research and improvement of the digitized material during the last years. We ended recently a two year European Regional Development Fund project and started another two year ERDF project.

→ **Digitalia (together with XAMK)**

- NLF is also involved in research consortium COMHIS that is funded by the Academy of Finland (2016–2019) and utilizes the newspaper and journal data in its research of historical changes of publicity in Finland.

Data improvement and new ways to use the data

- NLF has so far performed e.g. the following:
  - 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)
  - Several improvements for the Web interface (time-line, notebook property etc.)
  - Ground truth data for new optical character recognition
  - A new OCR process with Tesseract 3.04.01
  - Named Entity Recognition evaluation collections (two phases: initial trial and present with GT OCR data)

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 from pages of one newspaper (86 000 pages)
- Simple image classification, if feasible

Thank you!

Kimmo Kettunen
Mika Koistinen
Teemu Ruokolainen
The National Library of Finland, DH Projects

Digitalia Research Centre for Digital Information
XAMK
1640 THE NATIONAL LIBRARY OF FINLAND
UNIVERSITY OF HELSINKI Ruralia Institute
Programme for Sustainable Growth and Jobs
Leverage from the EU 2014–2020
European Union European Regional Development Fund