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

Kimmo Kettunen, Mika Koistinen and Teemu Ruokolainen
National Library of Finland, DH projects

National Library of Finland (NLF) has digitized historical newspapers, periodicals and ephemera published in Finland since the late 1990s. The present collection consists of about 11.8 million pages mainly in Finnish and Swedish. Out of these about 5.1 million pages are freely available on the web site digi.kansalliskirjasto.fi. The copyright restricted part of the collection can be used at six legal deposit libraries in different parts of Finland. The time period of the open collection is from 1771 to 1920. The last ten years, 1911–1920, were opened in February 2017.

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 current number of available data is probably already much bigger, as the national libraries have been working steadily with digitization both in Europe, Northern America and rest of the world.

Research and development

Besides producing the raw data all the time NLF has been involved in research and curation 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. NLF is also involved in research consortium COMHIS that is funded by the Academy of Finland (2016–2019) and utilizes the newspaper and periodical data in its research of historical changes of publicity in Finland.

So far we have focused on text material, and our main achievements have been the following: a thorough quality analysis on word level of the Finnish part of the textual data (Kettunen and Pääkkönen, 2016, Kettunen et al., 2016), evaluation of tools for named entity recognition and setting up of a NER evaluation collection (Kettunen et al., 2017) and an improved Optical Character Recognition framework for the data using Tesseract open source OCR engine (Koistinen et al., 2017). In March 2017 we published the newspaper and periodical text data as an open data collection on our web pages (Pääkkönen et al., 2016).

We have also started article extraction from the pages of one newspaper, and plan to do topic detection on the extracted articles in the future.

Newspapers and periodicals contain also lots of pictures: drawings, photographs and different graphs which are of interest to both researchers and lay persons. So far we can detect pictures on the pages of digi.kansalliskirjasto.fi, but a systematic way of both detecting the pictures and recognizing and classifying their content needs to be developed.

Acknowledgements

This work is funded by the European Regional Development Fund and the program Leverage from the EU 2014-2020.

References

Dunning, A. 2012. European Newspaper Survey Report. http://www.europeana-newspapers.eu/wp-content/uploads/2012/04/D4.1-Europeana-newspapers-survey-report.pdf.

Kettunen, K. and Pääkkönen, T. 2016. Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means. In Calzolari, N. et al. (eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) http://www.lrec-conf.org/proceedings/lrec2016/pdf/17_Paper.pdf.

Kettunen, K., Pääkkönen, T. and Koistinen, M. 2016. Kansalliskirjaston digitoitu historiallinen lehtiaineisto 1771–1910: sanatason laatu, kokoelmien käyttö ja laadun parantaminen. *Informaatiotutkimus* 3, 3–14. http://journal.fi/inf/article/view/59433.

Kettunen, K., Mäkelä, E., Ruokolainen, T., Kuokkala, J. and Löfberg, L. 2017. Old Content and Modern Tools – Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771–1910. *Digital Humanities Quarterly* (to appear).

Koistinen, M., Kettunen, K. and Pääkkönen, T. 2017. Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing. In Proceedings of Nodalida 2017 http://www.ep.liu.se/ecp/131/038/ecp17131038.pdf.

Pääkkönen, T., Kervinen, J., Nivala, A., Kettunen, K. & Mäkelä, E. 2016. Exporting Finnish Digitized Historical Newspaper Contents for Offline Use. *D-Lib Magazine*, July/August. http://www.dlib.org/dlib/july16/paakkonen/07paakkonen.html

The "State of the Art": A Comparative Analysis of Newspaper Digitization to Date. 2015. http://www.crl.edu/sites/default/files/d6/attachments/events/ICON_Report-State_of_Digitization_final.pdf.