National Library of Finland and Digitized Collections for Researchers

Tuula Pääkkönen, MSc ,Information Systems Specialist, National Library of Finland

The National Library of Finland aims to preserve and enable access to the cultural heritage materials of Finland, to researchers and all other users. The National Library has digitized all the newspapers until the year 1920 and most of the generic-purpose journals until the 1940’s. The material until the year 1920 is available via http://digi.kansalliskirjasto.fi for anyone in the world. The more recent digitized material is available at the six legal deposit libraries (National Library, Turku University Library, Jyväskylä University Library, Åbo Akademi University Library, Oulu University Library, University of Eastern Finland Library). We are also actively seeking ways to opening materials via contracts to researchers.

At the moment there are nearly 3 million newspaper pages until 1920. There are a bit over two million pages of journals, which are available via the public web system. Technical ephemera material, which includes e.g. price lists, brochures and catalogues exists in digi.kansalliskirjasto.fi with over 100.000 pages and more ephemera material can be found on doria.fi. The newspaper pages can be searched, viewed, downloaded as page image, text, XML or even data packages. The data packages contain all the pages in a custom XML format, which contains the metadata of the page, ALTO XML (which contains the layout analysis of the page as arrived from digitization) and page text as plain text.

Currently we are working towards improving the usability of the materials further in a few development and research projects. For example, we will have HAKA-authentication in digi.kansalliskirjasto.fi, which enables us to authenticate users from any university or other facility which uses HAKA. This enables us to open in-copyright materials. The aim in the other digital humanities projects is to improve OCR quality, find ways to utilize newspaper images, work onwards with article segmentation, to mention a few of our goals. Our current approach is to select one of the most used newspapers, namely Uusi Suometar, which will act as our proof-of-concept when improving further. Recently, for improving text quality our DH-projects have released a new ground truth material package, which enables comparing how well different OCR-correction methods would work with this particular material.