

OCR Ground Truth Package for Finnish Fraktur

Package contains 450 page images and ALTO XML files for each page, with the proofreading done by the Finnish native speakers.

Description

The pages of fraktur range from the year 1836 until 1910. The package can help in creating own postcorrection algorithms for OCR text recognition.

There is also an Excel file for all of the 471 903 words, which contains result given to the word by Tesseract and FineReader. If a tool hasn't found corresponding word, then the given cell is empty, so select the words in the Excel, which you need.

NB! The ground truth package does not contain the data for the 1918 due to copyright reasons.

User interface

Data downloads

<http://digi.kansalliskirjasto.fi/opendata>

API

-

License

[Terms of use](#) (in Finnish).

Content type

Pageimages, ALTO XML, metadata

Language

Finnish

Data status

Primary source

Size

450 pages

Update frequency

-

Relationships

-

External information

Article: Creating and using ground truth OCR sample data for Finnish historical newspapers and journals ([PDF file](#))

Acknowledgements

Digitalia project (<http://blogs.helsinki.fi/digitalia/>)

Leverage from
the EU
2014–2020



Contact information

kk-tutkijapalvelut@helsinki.fi

Star rating

-