# La verdad fundamental de una muestra de datos OCR de periódicos y revistas históricas finlandesas en la validación mejorada de datos de un proceso de doble OCR

Ground Truth OCR Sample Data of Finnish Historical Newspapers and Journals in Data Improvement Validation of a re-OCRing Process

Kimmo Kettunen, Mika Koistinen, Jukka Kervinen

*Liber quarterly*, ISSN 2213-056X, Vol. 30, n. 1, 2020

La Biblioteca Nacional de Finlandia (NLF) ha digitalizado desde finales de la década de 1990 periódicos, revistas y efímera históricas publicadas en Finlandia. La presente colección consiste en unos 16,51 millones de páginas principalmente en finés y sueco. De estos, alrededor de 7,64 millones de páginas son de acceso libre en la página web https://digi.kansalliskirjasto.fi/etusivu. La colección restringida por derechos de autor puede ser usada en cinco depósito bibliotecarios legales en diferentes partes de Finlandia. El periodo de tiempo de la colección abierta de de 1771 a 1929. Los últimos nueve años, 1921-1929, fueron abiertos en enero de 2018. Este artículo presenta brevemente la verdad fundamental de los datos de Optical Character Recognition de alrededor de 500.000 palabras compilados en la NLF para desarrollar un proceso de OCR mejorado para la colección finlandesa. Se trata la compilación de datos de manera general y se muestran los resultados del nuevo proceso OCR en comparación con el actual OCR, usando la verdad

fundamental de los datos cómo una evaluación comparativa. También se muestran con periódicos reales datos de 30 años y 109 millones de palabras que el proceso de doble OCR mejora la calidad de los datos OCRed.

https://www.liberquarterly.eu/articles/10.18352/lq.10322/

Traducción del resumen de la propia publicación