Tuesday, November 8th 2016

Breakthrough in archival access: Googling through archives within reach

Pilot project Full Automatic Archival Access completed

AMSTERDAM|THE HAGUE|MADRID|NIJMEGEN – The ability to google through records is within reach, concludes the final report of the project Full Automatic Archival Access (FAAA). This project studied the opportunities to use new digital technologies to make paper based archives searchable at document-level. Four out of five words were correctly recognized by OCR- and NER-software.

A small selection of the Central Archive of Justice (CABR, National Archives of the Netherlands) was used in the pilot. Project partners the Network of Dutch War Collections, Centre for Language and Speech Technology, National Archives of the Netherlands and IMPACT Centre of Competence are pleasantly surprised with the results.

Eighty-one percent of the words in the test-documents are correctly recognized by software. That means that it is possible to make typed or hybrid text documents with a standard layout automatically, digitally searchable with an acceptable error rate. A standard layout exists of straight lines, a regular ink density and clear contrast between text and background.

The FAAA-project consisted of two steps. First, the approximately one hundred documents from the CABR-archive have been made machine-readable with use of Optical Character Recognition (OCR)-software. Then, the quality of the OCR'ed-text was improved by using Named Entity Recognition (NER)-software. This software is able to select places, persons and organizations and correct them if necessary.

A leap forward in the accessibility of archives, which are currently mostly described on collection or sub-collection level and rarely accessible on document-level. Program director Network of Dutch War Collections Puck Huitsing: "The ability to make archives automatically digitally searchable offers many new opportunities for researchers. Historical collections can be questioned in a way that has never been possible in the paper world".

The project Full Automatic Archival Access was funded by Archief2020, BRAIN, VSBFonds, VFonds and the Ministry of Health, Welfare and Sport. The final and individual reports are published on the Network of Dutch War Collections-website: http://oorlogsbronnen.nl/volauto.

Organizations

The National Archives of the Netherlands in The Hague holds 125 kilometres of documents, photos and maps from the central government, and organizations and persons of national importance (past and present).

The Network of Dutch War Collections (NOB) is facilitated by NIOD Institute for War, Holocaust and Genocide Studies. It wants to enhance the use of sources from and about the Second World War in the Netherlands by making the scattered sources better findable and more usable.

The Centre for Language and Speech Technology (CLST) from the Radboud University Nijmegen aims to contribute to the development of language and speech technology. CLST is active in research, application development and consultancy.

The Impact Centre of Competence is a not for profit organisation, comprised of public and private institutions, with the mission to make the digitisation of historical printed text 'better, faster and cheaper'. It provides tools, services and facilities to further advance the state-of-the-art in the field of document imaging, language technology and the processing of historical text.