Contributors/Presenters:

Clemens Neudecker, Willem Jan Faber, Lotte Wilms
KB National Library of the Netherlands

Title:

Large scale refinement of digital historical newspapers with named entities recognition

Abstract:

Within the Europeana Newspapers project (www.europeana-newspapers.eu), full-text will be produced for over 10 million pages of digitised historical newspapers by applying Optical Character Recognition (OCR) and Optical Layout Recognition (OLR). In order to further increase the usability of the full-text, Named Entity Recognition (NER) is also applied to materials in Dutch, German and French language. The main aim of NER is to identify and classify entities such as persons, locations and organisations in the full-text in order to enhance the searchability, and to subsequently link them to online resource descriptions and authority files (e.g. DBPedia, VIAF). Therefore, the KB National Library of the Netherlands has been adapting an open source tool from Stanford University for training a state-of-the-art NER system specifically for Europeana Newspapers. Some of the main considerations in producing the software were scalability and the support for the standards and formats that are most widely used in newspaper digitisation such as METS and ALTO. Another important requirement was to retain information about the exact location of the named entities on a page throughout the refinement process, so they can be highlighted in a viewer. We will introduce the overall workflow for NER as implemented in Europeana Newspapers, discuss some design considerations as well as technical issues and lessons learned while building the software, and present first results from applying the system to historical newspaper content from three different languages.

Biographies:

Clemens Neudecker, M.A. Philosophy, Computer Science, Political Science, Technical Coordinator Research in the Innovation and Development department of the KB. He has been involved in numerous digitisation projects over more than a decade, previously at the Bavarian State Library. He has a particular interest in OCR and scalable digitisation workflows.

Willem Jan Faber, Research Programmer at the Innovation and Development department of the KB. Early adaptor of information technology, hacker and open source/Linux evangelist. He holds a community degree in information and communications technology and a Bachelor degree in the field of Computer Human Interaction design. In the past he also worked for internet companies XS4ALL and FOX-IT.

Lotte Wilms, KB Project Leader for the project European Newspapers Online. She has a BA in English Language and Culture and an MA in Medieval Studies from the University of Utrecht. She has worked at the KB since 2008 on various projects, such as the IMPACT project (www.impact-project.eu / www.digitisation.eu), Short-Title Catalogue Netherlands and the digitisation projects Staten-Generaal Digitaal and Early Dutch Books Online.