A HISTORIAN AND JOURNALIST TELLS HIS STORY

ANALYSING LANGUAGE

Ewoud Sanders, Dutch Language Historian and Journalist

Ewoud Sanders is best known for his weekly column WoordHoek ('Word Corner') in the newspaper NRC Handelsblad where he writes about the history of Dutch words and expressions. He is on a quest to improve digital access to printed Dutch language resources and his pamphlet Eerste Hulp Bij e-Onderzoek ("First Aid for e-Research") has been reprinted 16 times and distributed free of charge to students by several Dutch institutes of higher learning. In 2011, Google gave him a grant of $15,000 to help improve internet searching in the Netherlands.

HOW DO YOU USE TEXT AND DATA MINING?
I mostly use TDM for my work in the language column, which demands research of literary books. Furthermore, I have spent almost 15 years collecting large data sets of books written in Dutch by finding books that are already digitized and harvesting them. Within these 15 years I have learnt how to scan books properly and how to disclose them using OCR and index tools. Hence, I have indexed a collection of 250,000 books so far.

As already mentioned, I have been writing about the history of Dutch words and expressions. Using index tools I can search through my data collection in a much more advanced way than is possible on the internet. Using wildcards and performing proximity searches, for example, gives me an advantage in my research.

I also use my expert knowledge and skill in internet searching to prepare my publications. In 2009, I gave the 18th Bert van Selm memorial lecture on the latter topic, under the title *De reïncarnatie van het boek. In zeven stappen een eigen digitale bibliotheek* (The Reincarnation of the Book. Seven Steps Towards a Digital Library), which was published by Leiden University Press.

IS ACCESS TO CONTENT AN ISSUE FOR YOU?
It is legally possible to build such a database in my own environment for private use. I can do my research but I cannot share the content with other researchers. I help biographers and researchers up to the point that the legal limitations and copyright allow it. It is important for science to be able to share, but this possibility is very limited. Especially according to the Dutch legislation. One has to wait for 70 years after the author's death, to have full access and be able to share the content.

HOW DO YOU SEE TDM IN THE FUTURE?
Text and Data Mining needs to be an easier procedure without all these legal limitations. Lots of books are digitally born, for many years already, but there is no flow in the procedure of using them. More books should be available to researchers, especially literary. TDM should enable users further and help science move forward.

LEARN MORE
ewoudsanders.nl
wikipedia.org/wiki/Ewoud_Sanders

European Commission

ATHENA Research & Innovation Information Technologies

AGRO-KNOW

EMBL-EBI

gesis Leibniz Institute for the Social Sciences

grnet Networking Research and Education

MANCHESTER 1824 The University of Manchester

TECHNISCHE UNIVERSITÄT DARMSTADT

UNIVERSITY of STIRLING

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

EPFL ECOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

frontiers

INRA SCIENCE & IMPACT

LIBER LIBER

The Open University

UNIVERSITY OF AMSTERDAM

University of Glasgow

CREATe