DHBENELUX – LIÈGE 2019

THE ENHANCEMENT OF THE SEARCH-ABILITY OF OCRed TEXTS

HTR+, segmentation and meta-dating early modern ordinances

KB - NATIONAL LIBRARY OF THE NETHERLANDS (DEN HAAG)[1], GHENT UNIVERSITY[2], ERASMUS UNIVERSITY ROTTERDAM[3]
CHRISTEL ANNEMIEKE ROMEIN[123], SARA FLOOR VELDHOEN[1], AND MICHEL DE GRUIJTER[1]

Hypothesis: *When problems arose, small 'states' had to act swiftly. Governments may have adopted successful legislation from neighbouring areas. Hence 'entangled histories'.*
The project uses printed *books of ordinances* from the Low Countries (1500-1800s) to answer this.

Using Computer Technologies to read and follow Early Modern Political Legislation

Early modern ordinances were affixed to 'known places' and proclaimed by the city crier. (Paalhuis c. 1660).

Provincial governments decided to bundle the most important rules in books of ordinances (in Dutch: plakkaatboeken). These are not complete, but do give an overview of what government officials thought to be the most important rules that had been cast. Not all are digitised.

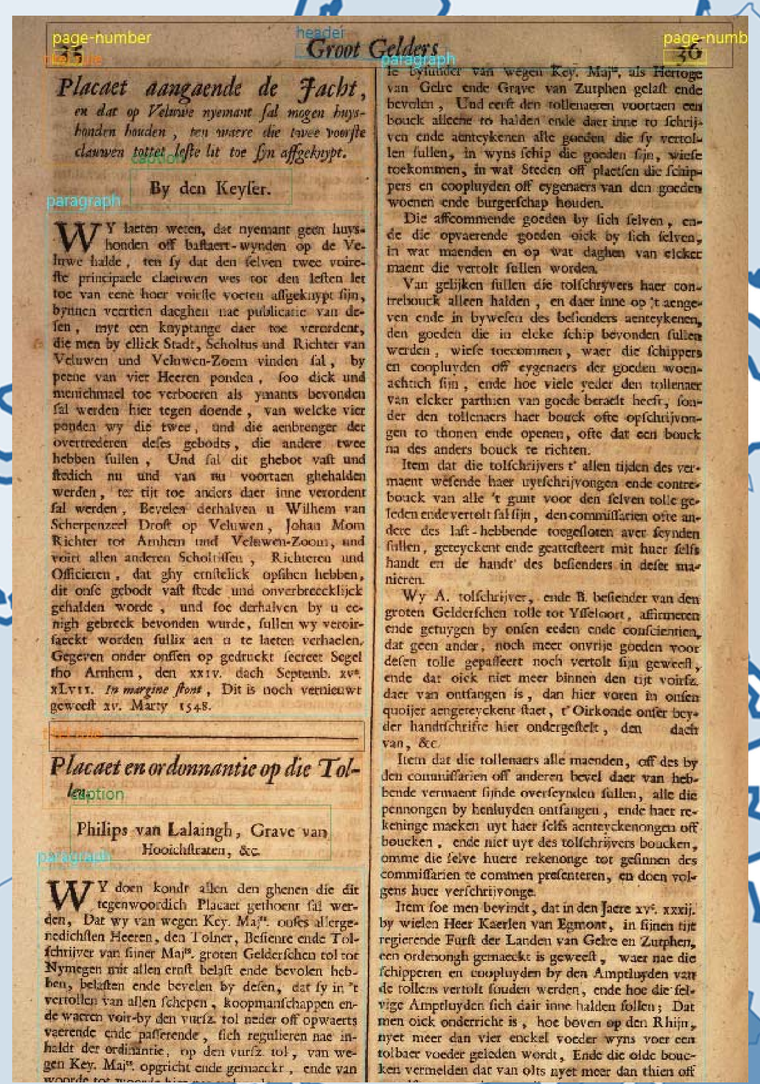
The set used comprises:

108 Books:

- font:
  - roman 88
  - gothic 20
- language:
  - Dutch 67
  - French 26
  - Latin 1
  - Mixed 14
- pages: 75.372
- words: ± 550 mln.

Steps within the project:

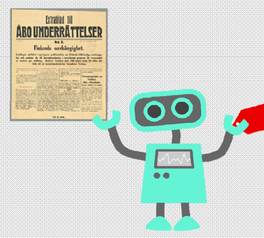
1. Within 'Entangled Histories' we apply the P2PaLA-tool of the University of Valencia to recognise sections in the text.



2. Creating "Handwritten Text Recognition-models" with Transkribus.
- language (Dutch/French)
- font (roman/gothic/variations)
- character Error Rate <5%

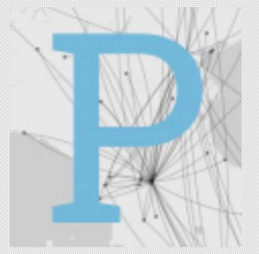
This data we can export as XML/TEI files (but also as PDF/A or Word-files).

3. Machine-learning categorisation through the use of ANNIF through the GhentHPC:
- hierarchical categories created by the MPI für europäische Rechtsgeschichte (1800 in total)
- per individual law
- based on manually labelled data (pre-trained)



Annif (annif.org)

Results:
- searchable data.
- visualisations through plotting
- on maps (Palladio for example).
- data will be connected through Linked Data and URI's with the dataset in Frankfurt.



e.g. Palladio - Stanford

Project-team

Michel de Gruijter – Project Advisor
Annemieke Romein – Primary Investigator
Sara Veldhoen – Research Software Engineer

Contact

Annemieke.Romein@ugent.be
https://research.flw.ugent.be/nl/annemieke.romein
www.caromein.nl

@caromein.nl
Christel Annemieke Romein

GHENT UNIVERSITY · Erasmus University Rotterdam · KB national library of the netherlands