Exploring text mining techniques to structure a digitised catalogue

Karen Goes [1], Sara Veldhoen [2] & Steven Claeyssens [2]
kwmgoes@gmail.com, {sara.veldhoen; steven.claeyssens}@kb.nl
[1] VU University Amsterdam; [2] KB, National Library of the Netherlands

Introduction

This research is conducted as part of an internship at the National Library of The Netherlands. The National Library collects all books that are published in the Netherlands, this has actively been done since 1974.

The Brinkman Catalogue of Books is part of the collection and the subject of this research. The catalogue lists all the books published in the Netherlands within a specific time frame. It is available in monthly and yearly volumes, as well as volumes covering multiple years. These volumes are part of the national bibliography, and are currently only available as unstructured data, on paper and partially digitised.

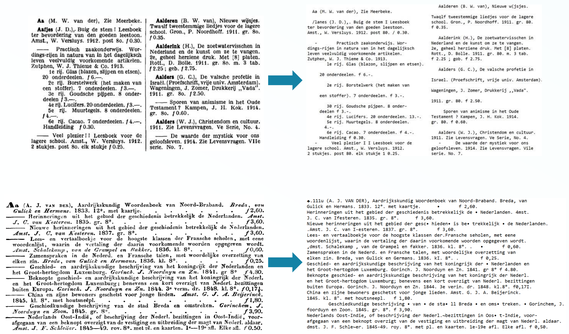
This research uses the digitised catalogue volumes to extract the data from them and turn them into structured data, that is machine-readable and searchable. To accomplish this the use of multiple text mining techniques is explored to determine the best method for this task.

Using the structured data the National Library can make a comparison with their own collection to identify books that are missing from their collection, especially from before 1974.

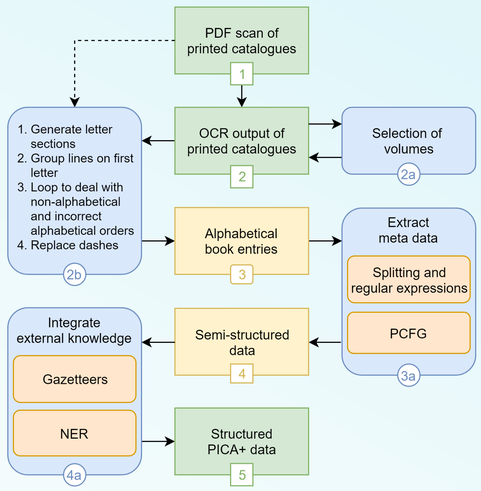
Materials

31 Brinkman catalogue volumes, originally published between 1833 and 1980. Two material types are available per catalogue volume:

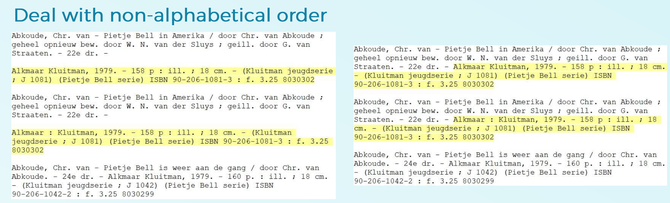
- PDF scan
- OCR output



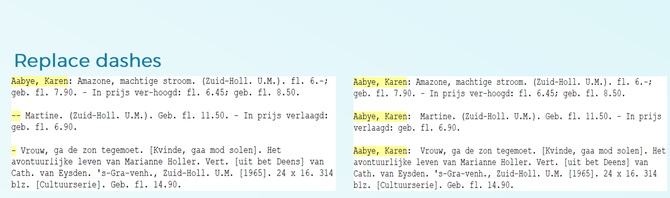
Methodology



Deal with non-alphabetical order



Replace dashes



Extracted meta data

Author	Title	City	Publisher	Year	Pages	Size	ISBN	Price	Product number
Unknown	A Paris	Zutphen	Thieme	1978	25	30 cm	9003285004	f. 5.20	7905126
Unknown	A Paris	Zutphen	Thieme	1979	25	30 cm	9003285004	f. 5.20	7938217
Unknown	Unknown	A +	aardrijkskunde.	Unknown	15	22 cm	902571112X	f. 3.25	7924032
Aa, A. J. van der	Aardrijkskundig woordenboek der Nederlanden	Zaltbommel	Europese Bibliotheek.	Unknown	136	18 cm	N/A	f. 55	7928260
Aafjes, Bertus	Kleine Isar, de vierde koning : halleluja, de wonde	Amsterdam	Meulenhoff	1979	217	14 x 21 cm	9029012323	f. 25	7956335
Aafjes, Bertus	Het koningsgraf : honderd en een sonnetten	Bussum	Agathon	1979	119	19 cm	9026956436	f. 9.90	7931389
Aafjes, Bertus	Een ladder tegen een wolk	Amsterdam	Querido	1979	158	19 cm	902149468X	f. 6.20	7926190
Aafjes, Bertus	Limburg, dierbaar oord	Amsterdam	Meulenhoff	1979	Unknown	Unknown	9029007915	f. 26.50	7927311

Evaluation

❶ Manually created evaluation data

For a total of 100 bibliographical entries the meta data is manually copied down to create evaluation data. Entries from random pages in a volume are copied until the 100 entries are reached. This is done for all volumes from 1850 to 1971.

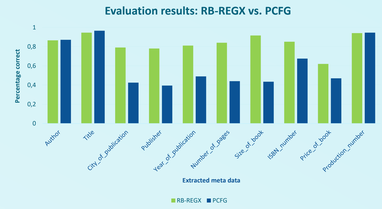
Guidelines are created to ensure that the process of copying down the meta data is done in the same manner across all the volumes and entries.

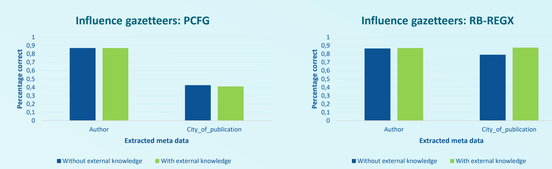
❷ Digitally available evaluation data

For the volumes from 1975 to 1980, the National Library has provided data that can be used as evaluation data. This data can be linked to the extracted data using the Brinkman production number, which can be found in both the catalogues and the library data.

Again, 100 bibliographical entries are used for the evaluation taken at random. With as selection requirement a Brinkman production number that can be linked to the library data, which is already digitally available.

Evaluation results





Conclusion

The evaluation results show that the RB-REGX system is the best technique as it is able to extract meta data from all the entries and with a high accuracy.

Across all the volumes the lowest extraction percentage is for the ISBN with 27%, however, not all entries have an ISBN number.

The RB-REGX technique is more efficient to implement, and it performs better than the PCFG.

Future recommendations

- Redo the OCR process to minimise the loss of volumes and simplify the formation of bibliographical entries.
- Use the output of the current research as annotated data for a machine learning system.
- Process other parts of the catalogue such as the bibliographical entries sorted by topic.