Refugee Policy in 1970s Archives

Sara Cosemans (KU Leuven) Philip Grant (independent researcher)
Ratan Sebastian (Leibniz University, Hannover)
Marc Allassonière-Tang (Museum of Natural History, Paris)

Research questions

Historical

- In the 1970s, why did **refugee resettlement** become the favoured solution for the protection of Ugandan Asian, Chilean and Vietnamese refugees?
- How did information and policy ideas travel: ”horizontally” between states and organizations, and ”vertically” through time?

Methodological

- How well can off-the-shelf programs deal with messy paper-born typewritten archival documents?
- What is the optimal approach for topic modelling based on imperfect data?

Starting out

- 23,818 scanned typewritten pages
- 8,645 digitised cables

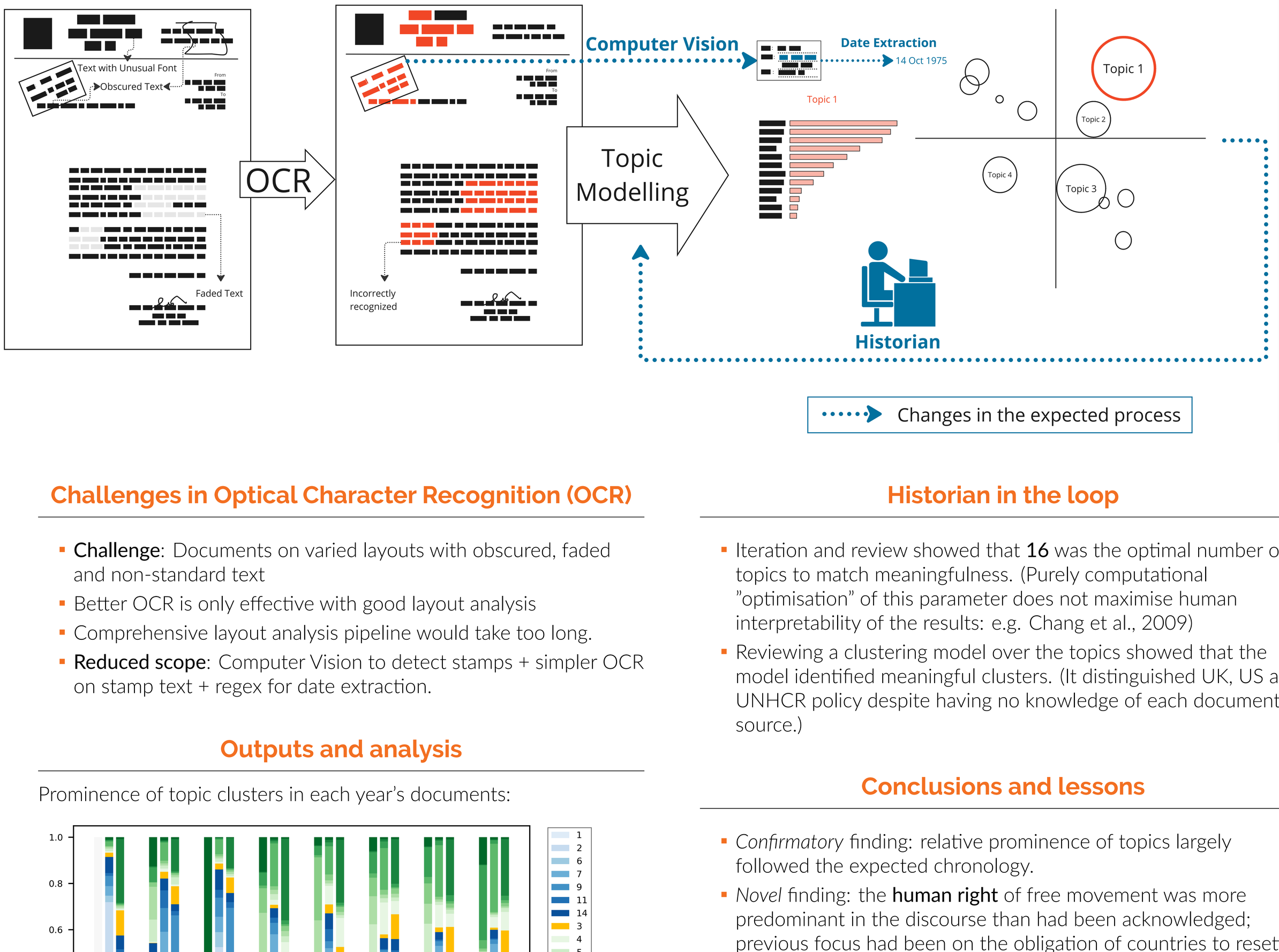
We looked at digital techniques to:

- Extract text and metadata (document dates) from the scanned pages
- Give ”distant reading” insights into the content of the archives

We expected the scanned documents to be challenging:

- Dates often appeared as literal rubber stamps
- Paper and print quality was often poor - problematic for OCR

The process and how we adapted it



Challenges in Optical Character Recognition (OCR)

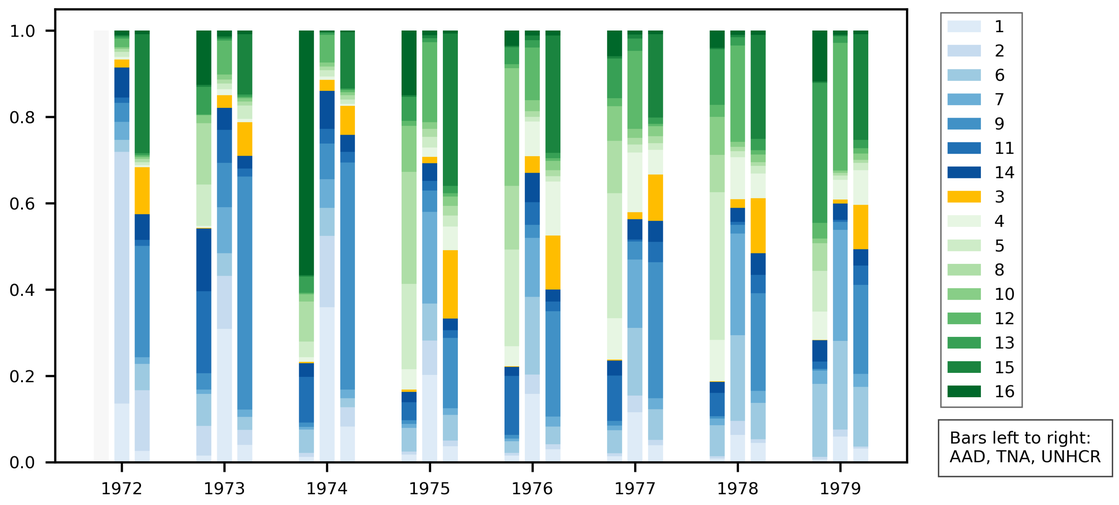
- **Challenge**: Documents on varied layouts with obscured, faded and non-standard text
- Better OCR is only effective with good layout analysis
- Comprehensive layout analysis pipeline would take too long.
- **Reduced scope**: Computer Vision to detect stamps + simpler OCR on stamp text + regex for date extraction.

Historian in the loop

- Iteration and review showed that **16** was the optimal number of topics to match meaningfulness. (Purely computational ”optimisation” of this parameter does not maximise human interpretability of the results: e.g. Chang et al., 2009)
- Reviewing a clustering model over the topics showed that the model identified meaningful clusters. (It distinguished UK, US and UNHCR policy despite having no knowledge of each document’s source.)

Outputs and analysis

Prominence of topic clusters in each year’s documents:



Blue: UK policy; **Green**: US policy; **Orange**: UNHCR

Conclusions and lessons

- *Confirmatory* finding: relative prominence of topics largely followed the expected chronology.
- *Novel* finding: the **human right** of free movement was more predominant in the discourse than had been acknowledged; previous focus had been on the obligation of countries to resettle refugees.
- Even with **very imperfect scanning and OCR**, topic modelling gives useful insights: a ”macro” view, and a pointer to useful ”micro” views through close reading.

Digital Humanities Benelux, 6 June 2024