Transcrire des doléances au XXI[e] siècle : le cas des Cahiers Citoyens issus du Grand Débat National

Transcribing Grievances in the 21st Century: The Case of the "Cahiers Citoyens" from the Great National Debate

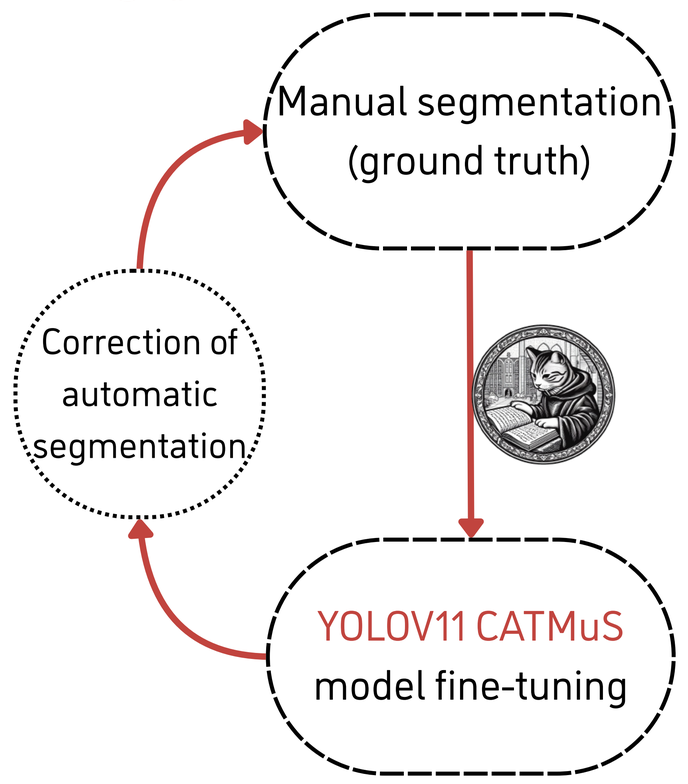
The "Cahiers Citoyens" refer to notebooks made available in more than 20,000 municipalities around France to collect citizens' grievances between November 2018 and February 2019, in response to the Yellow Vests protests and following the government's launch of the Great National Debate (GDN). They comprise approximately 450,000 pages of handwritten and printed contributions.

As a proof of concept, this project focuses on 5% of the dataset, with the aim of publishing the entire corpus online with complete and accurate transcriptions of the text, produced through Automated Text Recognition (ATR), along with the model. Given the data's identifying and sensitive nature, a thorough anonymisation process is necessary to meet Open Science standards.

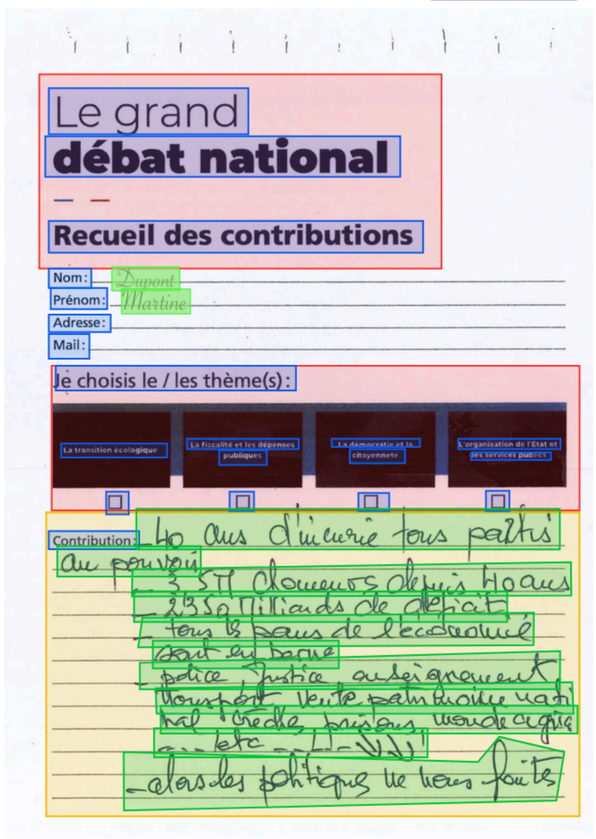
Segmentation

Challenges:

- Varied formats: small or large-grid notebooks, blank pages, forms mixing typewritten and handwritten content;
- Non-linear writing: rarely straight text, with skewed and irregular baselines;
- No preexisting page and element typology suited for this corpus.



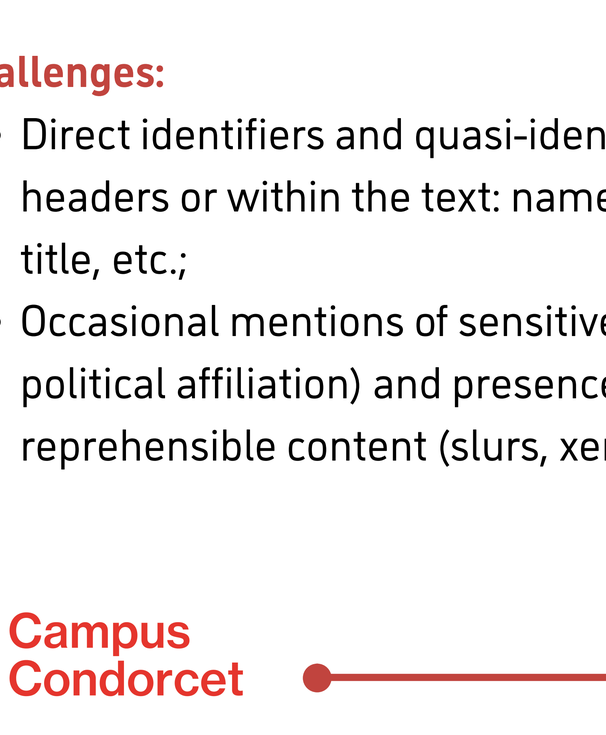
Working on Arkindex

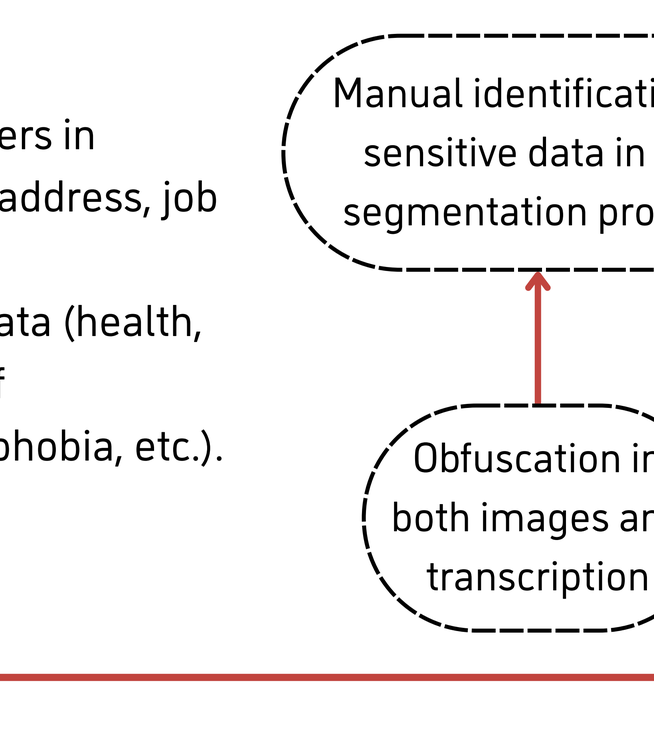


Transcription

Challenges:

- Working on digitized photocopies of the notebooks, with discrepancies in image quality;
- No existing public dataset on contemporary French handwriting;
- Frequent changes in script (~250,000);
- Spelling and grammatical errors;
- Overuse of jargon and acronyms.





Anonymisation

Challenges:

- Direct identifiers and quasi-identifiers in headers or within the text: names, address, job title, etc.;
- Occasional mentions of sensitive data (health, political affiliation) and presence of reprehensible content (slurs, xenophobia, etc.).

Campus Condorcet

Julie Lenoir, Léa Maronet, Alyx Taounza-Jeminet

With the support of the Ministère de l'Enseignement Supérieur et la Recherche et de l'Espace

With the help of Vincent Desmet, Salomé Ferber, Garance Heloir, Ismène Lemaire, Mélite Hestault, LeThuyTien Nguyen, Anaëlle Porteau, Eloïse Ripert-Fabre and Cyprien Valora.

Huma-Num IR*