Towards a datafication of Antwerp street life? Co-creating a dataset of 100.000+ pages of handwritten police reports (1876-1945)

Lefranc, Lith
lith.lefranc@uantwerpen.be
University of Antwerp, Belgium

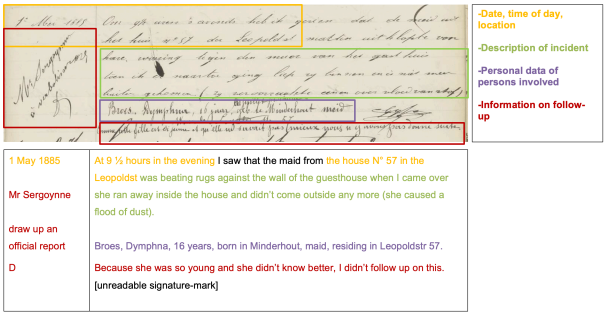
Since the 1960s, when the so-called 'new social historians' started questioning mainstream historical narratives and experiences, new historical sources such as police and judicial sources came under scrutiny, putting the spotlight on social groups and 'people without history' who have left almost no paper trail of their own (Thompson 1963; Wolf 1982). These historical documents have been valuable for a myriad of approaches which are not necessarily crime-related: working-class history, the history of mentalities, gender and sexuality history, youth and family history, *Alltagsgeschichte*, and so on. Up until today, qualitative methodological approaches have dominated this research area, emphasising thick description and case-based research. Only recently, social historians have been discovering the potential of data-driven approaches towards criminal records (e.g. Van den Heuvel 2020; Saygi / Yasunaga 2021). One of the most straightforward explanations for this lag is the lack of machine-readable data. Nineteenth-century bureaucratisation processes have left historians of the modern period with an abundance of archival material, but only the tip of the iceberg is digitally available, let alone searchable. Grand digitisation initiatives such as Old Bailey Online remain scarce because they require a lot of time, money, and labour (Hitchcock et al. 2012). In response, crowdsourcing platforms have been popping up like mushrooms, but these are often ill-suited for researchers with limited means such as PhDs.

Over the past few years, the technique of Handwritten Text Recognition (HTR) – made accessible through e.g. Transkribus and eScriptorium – improved greatly, which broke new ground for small-scale research projects (Muehlberger et al. 2019; Kiessling et al. 2019). In this poster, I will present a dataset of approximately 100.000 HTRed pages of local police reports from the city of Antwerp (1876-1945). Apart from the dataset specifics, I will show the challenges for HTR-training, which relied on both gold standard expert transcriptions, and semi-gold standard material provided by undergraduate students. Both collections will be scrutinised, and solutions will be examined for bypassing noisy training data. I hope that this presentation not only encourages qualitative social historians to test their hypotheses on 'big data', but also inspires them to step by step unlock more underexplored paper archives for data-driven research.

Dataset specifics

Incident books

From the 1860s onwards Antwerp police officers were obliged to write a daily report after patrolling their respective neighbourhoods in the so-called 'incident books'. One of the major historiographical benefits is the fact that these reports not only document prosecuted crimes, but non-prosecuted misdemeanours and other irregularities as well. Therefore, they give a very extensive account of all sorts of happenings on the Antwerp streets, such as neighbourhood quarrels, monkey tricks, lost children… The serial character of the incident books ensures a consistent registration of who (gender, profession, age, place of residence and birth) was when (date and time of day) and where (address) doing what (detailed description of incident) (see image below).



Space and time

The dataset consists of 326 books which date from 1876 to 1945 and are located within 11 different police districts. The data is not spread evenly throughout space and time. Most pages contain police reports dedicated to the third and the sixth district (predominantly bourgeois districts), and the page number increases during the 1930s and the Second World War (a period of increasing police surveillance).

HTR-training results

The incident books were very challenging for HTR-training because they contain numerous different handwritings and are written in French as well as Dutch. Furthermore, the difficult and irregular layout of the reports made segmentation very time-consuming. I have worked with two different training sets: gold standard expert transcriptions, and semi-gold standard material provided by undergraduates. The former consists of 326 pages selected randomly from each incident book and the latter contains several samples (in total 3000 pages).

Despite the imperfect undergraduate work, the models trained on all the material performed best, but unfortunately only led up to a (still unsatisfactory) character error rate (CER) of 9.3%. After systematically analysing the word and character errors of this model, I managed to improve the CER up to 6.3% by eliminating skewed text and signature marks. Finally, I switched to a more relaxed evaluation of the model by lowercasing all text and excluding interpunction and redundant spacing characters.

The fully HTRed dataset consists of about 30,5 million words and will soon be published in open-access format.

Bibliography

Kiessling, Benjamin et al. (2019): “eScriptorium: An Open Source Platform for Historical Document Analysis”, in: International Conference on Document Analysis and Recognition Workshops (ICDARW), 2: 19–24.

Hitchcock, Tim et al. (2012): The Old Bailey Proceedings Online, 1674-1913, https://www.oldbaileyonline.org [4.11.2022];

Muehlberger, Guenter et al. (2019): “Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study”, I,: Journal of Documentation 75, 5: 954–976.

Saygi, Gamze / **Yasunaga, Marie** (2021): “The digital urban experience of a lost city using mixed methods to depict the historical street life of Edo/Tokyo”, in: Magazén 2, 2: 193-224.

Thompson, Edward Palmer (1963): The Making of the English Working Class. London: Victor Gollancz Ltd.

van den Heuvel, D. et al. (2020): “Capturing gendered mobility and street use in the historical city: a new methodological approach”, in: Cultural and Social History 17, 4: 515–536.

Wolf, Eric (1982): Europe and the People without History. Berkeley: University of California Press.