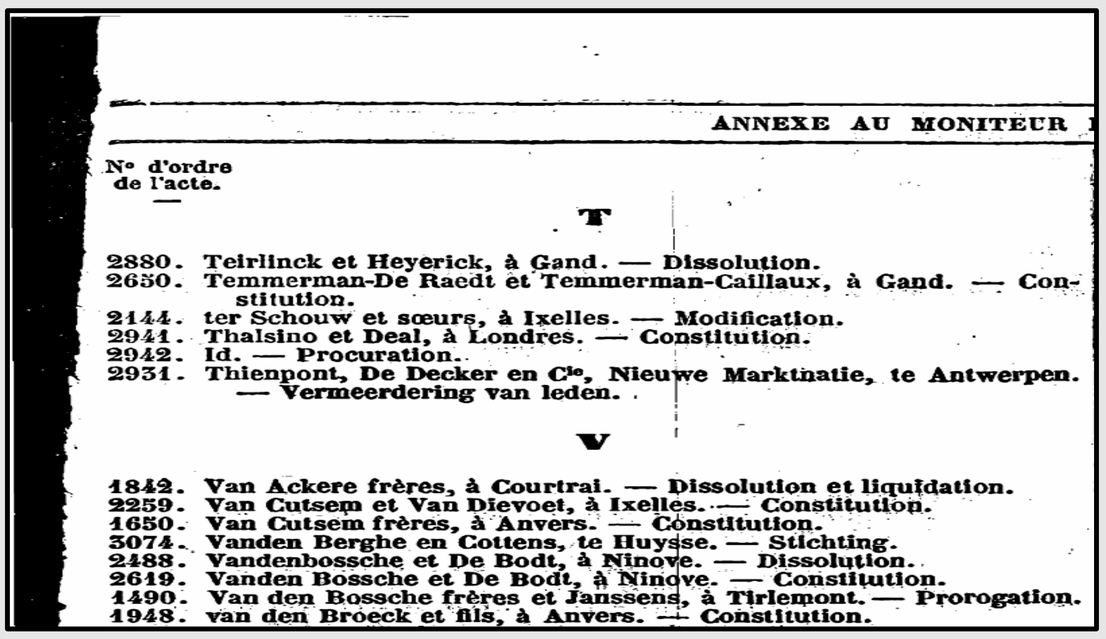
FACULTY OF ARTS AND PHILOSOPHY

GHENT CENTRE FOR DIGITAL HUMANITIES (GHENTCDH)

Bas Vercruysse, Vincent Ducatteeuw, Robrecht Declercq, Julie Birkholz and Christophe Verbruggen

AN OCR-WORKFLOW FOR SEMI-STRUCTURED HISTORICAL SOURCES

How to make economic data searchable?



ANNEXE AU MONITEUR

N° d'ordre de l'acte.

T

2880. Teirlinck et Heyerick, à Gand. — Dissolution.
2650. Temmerman-De Raedt et Temmerman-Caillaux, à Gand. — Constitution.
2144. ter Schouw et sœurs, à Ixelles. — Modification.
2941. Thalsino et Deal, à Londres. — Constitution.
2942. Id. — Procuration.
2931. Thienpont, De Decker en Cie, Nieuwe Marktnatie, te Antwerpen. — Vermeerdering van leden.

V

1842. Van Ackere frères, à Courtrai. — Dissolution et liquidation.
2259. Van Cutsem et Van Dievoet, à Ixelles. — Constitution.
1650. Van Cutsem frères, à Anvers. — Constitution.
3074. Vanden Berghe en Cottens, te Huysse. — Stichting.
2488. Vandenbossche et De Bodt, à Ninove. — Dissolution.
2619. Vanden Bossche et De Bodt, à Ninove. — Constitution.
1490. Van den Bossche frères et Janssens, à Tirlemont. — Prorogation.
1948. van den Broeck et fils, à Anvers. — Constitution.

2144. ter Schouw et sœurs, à Ixelles. — Modification.

From index to insight

The main objective of this workflow is to accurately extract and parse textual data from index lists into a tabular format, making the data searchable. Key challenges include detecting text lines accurately **(1)**, obtaining precise OCR results **(2)**, and correctly categorizing each element within an index entry **(3)**.

Each entry consists of four elements: a record number, company or organization names (including alternative names), location, and event. Though entries share a consistent structure, the length of each element varies significantly, and entries can span multiple lines. Traditional methods like regular expressions lack the flexibility needed for this complex structure. Our workflow utilizes machine learning and advanced AI models for efficient detection, extraction, and parsing of these entries.

2931. Thienpont, De Decker en Cie, Nieuwe Marktnatie, te Antwerpen.
— Vermeerdering van leden.

1) Text line detection

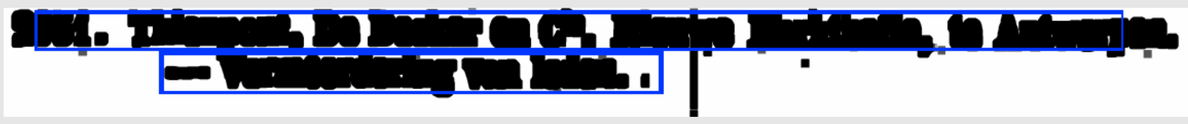
To detect lines on a page, we combined an object detector—typically used for recognizing animals like cats or dogs—and traditional image manipulation to extract index entries from our source material.

We fine-tuned a model from the You Only Look Once (YOLO) V11 family to detect lines of text. Each line is converted into a black blob for improved detection accuracy. However, our object detector sometimes misses the exact beginning and end of a sentence.

In the next step, we refine the detection coordinates using the full image width to capture a complete line. We then perform a second beginning-of-sentence (BOS) detection to ensure completeness.

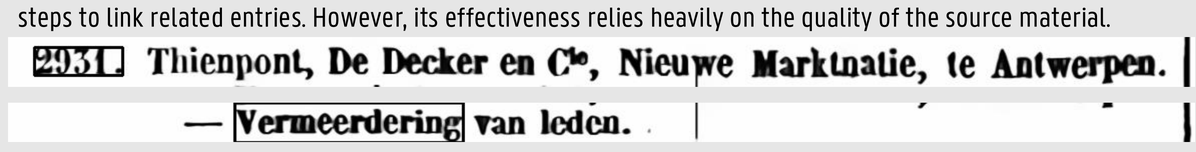
Yolov11 Detection

Our testing showed that 99.6% of sentences are detected (Recall) using this method, but only 75% of sentences are considered fully detected (maP50-95).



Refined BOS detection

This method leverages image manipulation techniques to locate the first word in a sentence, which helps in later steps to link related entries. However, its effectiveness relies heavily on the quality of the source material.



2931. Thienpont, De Decker en Cie, Nieuwe Marktnatie, te Antwerpen.
— Vermeerdering van leden.

Line-by-line Character Recognition

Although vision language models can process an entire image at once, this approach tends to increase the risk of 'hallucinations', sometimes observed with generative AI-models. We found that a line-by-line OCR method delivers significantly better results, though it is more computationally intensive.

OCR output
2144, ter Schouw et sœurs, à Ixelles. – Modification.
2931, Thienpont, De Decker en Cie , Nieuwe Marktnatie, te Antwerpen
- Vermeerdering van leden,

2) OCR using Qwen-2-VL-7B

The introduction of transformer models (like **ChatGPT**) and its visual counterpart, new developments in OCR have been quickly made. The combination of both types in so-called **vision language** models, seem to have increased the power of these transformer models even further.

Qwen-2-VL is a series of open-weights vision language models that can understand images based on a prompt. We tested 124 text line images with the standard (not fine-tuned) Qwen-2-VL-7B model. This model has an average Character Error Rate (CER) of 1% and Word Error Rate (WER) of 6%. Even more impressive is that the quality of the data is almost always to blame for the mistakes.

3) Using Conditional Random Fields for data parsing

Conditional Random Fields (CRF) is a statistical modelling method which can be used for pattern recognition. We leverage this capability for splitting our data (that follows a predefined structure) in certain classes. We use the coordinates from the bounding boxes of the refined BOS detection to combine text lines if required. This approach yields great results and is much more robust than regular expressions.

A searchable database

The result is a searchable database containing each entry's record number, name, location, and event. Verified by researchers, this streamlined workflow significantly accelerates data processing. In the future, the data will be integrated into a larger database accessible to researchers.

RecordID	Name	Location	Event
2144	ter Schouw et sœurs	à Ixelles	Modification
2931	Thienpont, De Decker en Cie, Nieuwe Marktnatie	te Antwerpen	Vermeerdering van leden

The BelHisFirm Project

Until now, most long-term economic data for Flanders and Belgium has been available only at the macro level, limiting insights into socio-economic changes over time. To delve deeper, we need micro-level data. The *Moniteur Belge/Belgisch Staatsblad*, which has documented Belgian government publications since the 19th century, holds a wealth of untapped micro-economic information in its appendices—covering company profiles, financial records, directors' and shareholders' personal details, profits and losses, and inter-company relationships.

"BelHisFirm: Long-Term Business Data for the Social Sciences" aims to make this valuable data accessible through a searchable database and visualization tools. By applying Optical Character Recognition (OCR) and page segmentation to over 2 million pages, the project will digitize and extract essential information, starting with indices listing names, events, and reference numbers.

This large-scale infrastructure (FWO), a collaboration between the University of Antwerp and Ghent University under the supervision of Christophe Verbruggen, Julie Birkholz, and Robrecht Declercq, will unlock vital micro-economic data. It will enable pioneering research into corporate finance, wealth inequality, crises, elite networks, and the economic impact of (de)colonization.

Contact

bas.vercruysse@ugent.be
https://www.ghentcdh.ugent.be/team/bas-vercruysse

Bas Vercruysse

https://github.com/GhentCDH

GHENT UNIVERSITY