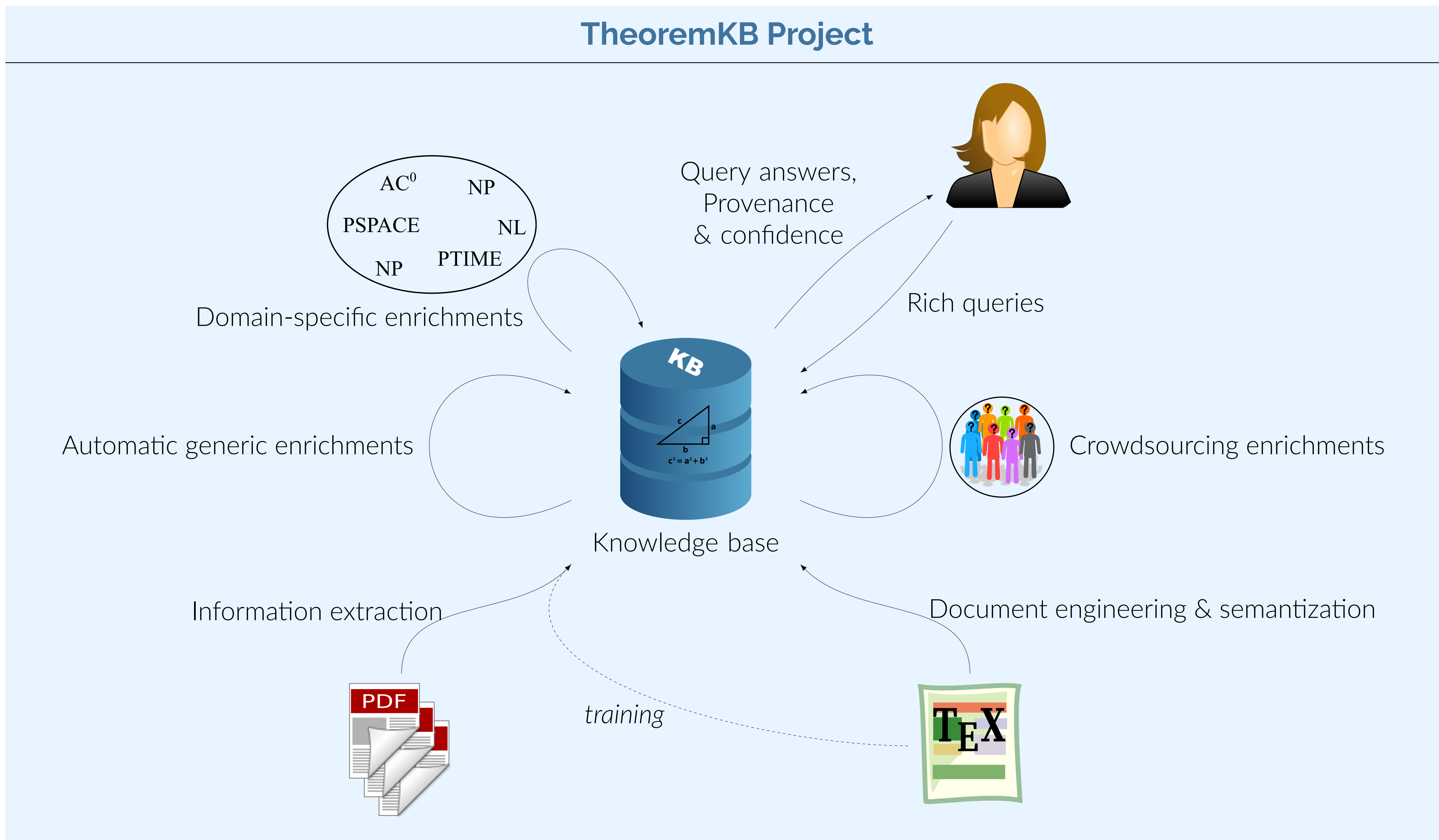


# First Steps in Building a Knowledge Base of Mathematical Results

Shrey Mishra Yacine Brihouché Théo Delemazure Antoine Gauquier Pierre Senellart



## First steps: Extraction + Linking

## Linking: Problem & Methodology [2, 1]

PDF Dataset → Extracted results → Result graph

The diagram shows the extraction process. On the left, a PDF snippet from 'The logarithm remains of Taylor's series, diverging' is shown. In the middle, 'Extracted results' are listed as numbered lemmas and theorems. On the right, a 'Result graph' shows these results as nodes connected by arrows, representing dependencies between mathematical statements.

Within the proof of a theorem, identify which result is used (and therefore which result the theorem depends on): from which paper does the result come from? which specific result from that paper is used?

The diagram illustrates the linking process. It starts with a 'PDF file' and an 'XML file'. The XML file is processed by 'Semantic Scholar + GROBID' to produce 'Citations'. The XML file is also processed to produce 'Results' (Corollary 4.3, Theorem 4.4, Theorem 4.5, Definition 4.6) and 'References' (Corollary 4.3 - Theorem 4.5, [8]). The final output is a 'Graph of mathematical results'.

## Extraction: Problem & Methodology [4]

Classify whether a paragraph of text is part of a mathematical statement (theorem, definition, etc.), part of a proof, or neither (basic text).

We train (deep learning) classifiers from an automatically labeled dataset from arXiv:

- Font (LSTM) Use the sequence of fonts assigned to each character of a paragraph
- Vision (EfficientNet) Use the bitmap rendering of the PDF
- Text (custom-trained BERT-like LM) Use the text extracted from the PDF
- Multimodal (GMU) Integration of features from 3 modalities
- Sequence model On each unimodal and multimodal model, also take into account the sequence of labels

## Extraction: Preliminary Results

Modality	Seq. approach	#Batches	#Params (M)	Accuracy (%)	Mean F <sub>1</sub> (%)
Dummy	—	—	—	59.41	24.85
Line-based [5]	—	—	110	57.31	55.71
Font	—	11	2	64.93	45.48
	CRF	11+1	2	71.50	64.51
Vision	—	9	53	69.44	60.33
	CRF	9+1	53	74.63	70.82
Text	—	20	124	76.45	72.33
	CRF	20+1	124	83.10	80.99
Multimodal	—	10	185	76.86	73.87
	CRF	10+1	185	<b>84.19</b>	<b>82.91</b>

The diagram shows the linking process from a PDF file to an XML file, then to a graph of mathematical results. It includes a sample PDF snippet and a corresponding graph of mathematical results. The graph shows nodes representing mathematical results and their dependencies. The diagram also includes a list of references.

## References

- Yacine Brihouché. TheoremKB : une base de connaissance des résultats mathématiques. Master's thesis, Paris IX Dauphine, September 2022.
- Theo Delemazure. A Knowledge Base of Mathematical Results. Master's thesis, Ecole Normale Supérieure (ENS), September 2020.
- Shrey Mishra, Yacine Brihouché, Théo Delemazure, Antoine Gauquier, and Pierre Senellart. First Steps in Building a Knowledge Base of Mathematical Results. In Proc. SDP, Bangkok, Thailand, August 2024.
- Shrey Mishra, Antoine Gauquier, and Pierre Senellart. Multimodal machine learning for extraction of theorems and proofs in the scientific literature. CoRR, abs/2307.09047, 2023.
- Shrey Mishra, Lucas Pluvinau, and Pierre Senellart. Towards extraction of theorems and proofs in scholarly articles. In Patrick Healy, Mihai Bilauca, and Alexandra Bonnici, editors, DocEng '21: ACM Symposium on Document Engineering 2021, Limerick, Ireland, August 24-27, 2021, pages 25:1-25:4. ACM, 2021.