Institut für Informationssysteme
Technische Universität Braunschweig

A Toolbox for the Nearly-Unsupervised Construction of Digital Library Knowledge Graphs

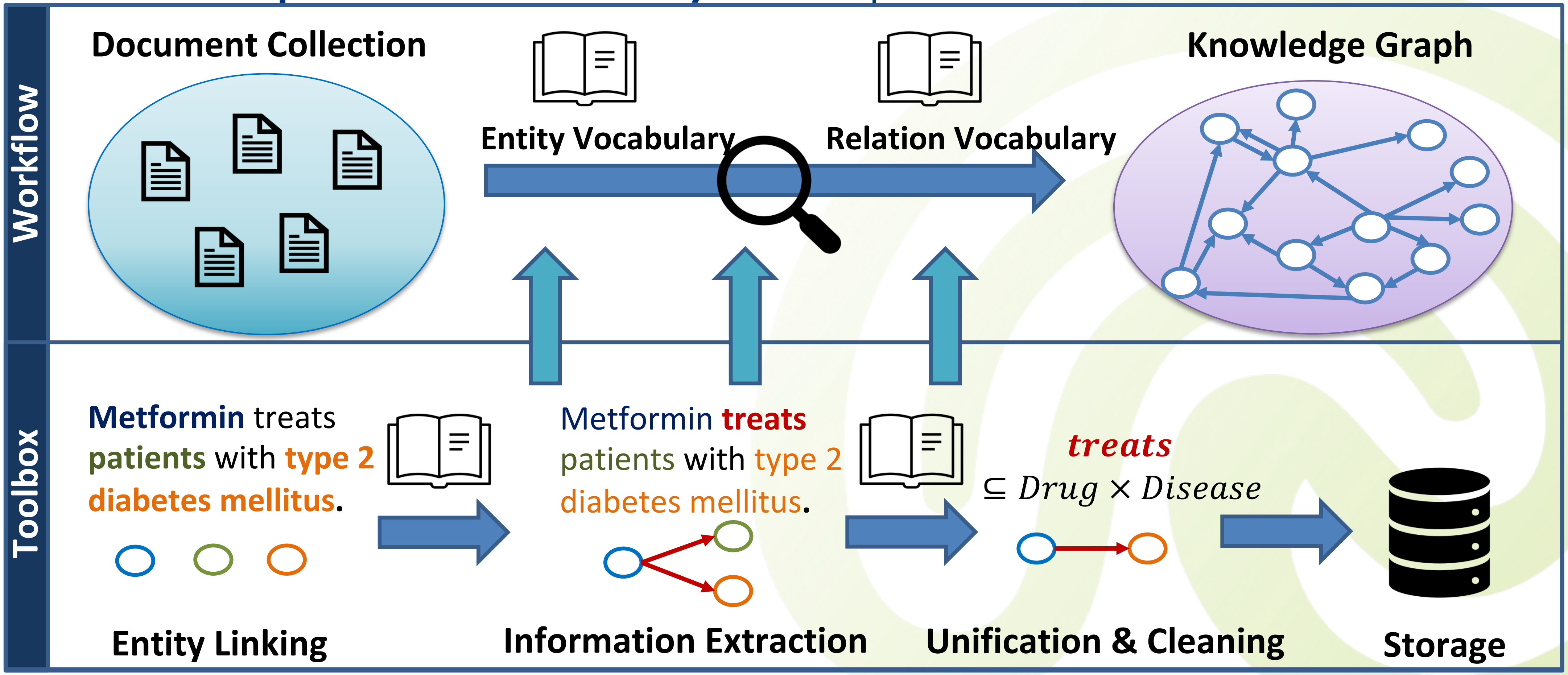
at JCDL2021

Hermann Kroll, Jan Pirklbauer and Wolf-Tilo Balke

Institut für Informationssysteme
Technische Universität Braunschweig

Toolbox Overview

- Available at:
  - https://github.com/HermannKroll/KGExtractionToolbox
  - Shared as **Open Source**, written in **Python** and published with an **MIT license**



Entity Linking & Recognition

- Dictionary-based **Entity Linker**:
  - Fast and reliable entity linking against a pre-known entity vocabulary
  - Not as precise as domain-specific entity linking
- Domain-specific tool integration:
  - As a demo, we integrated two biomedical tools
- Stanford Stanza (**Named Entity Recognition**):
  - Detects entities such as persons, events, dates, etc. (18 types)
  - Does not deliver unique entity ids

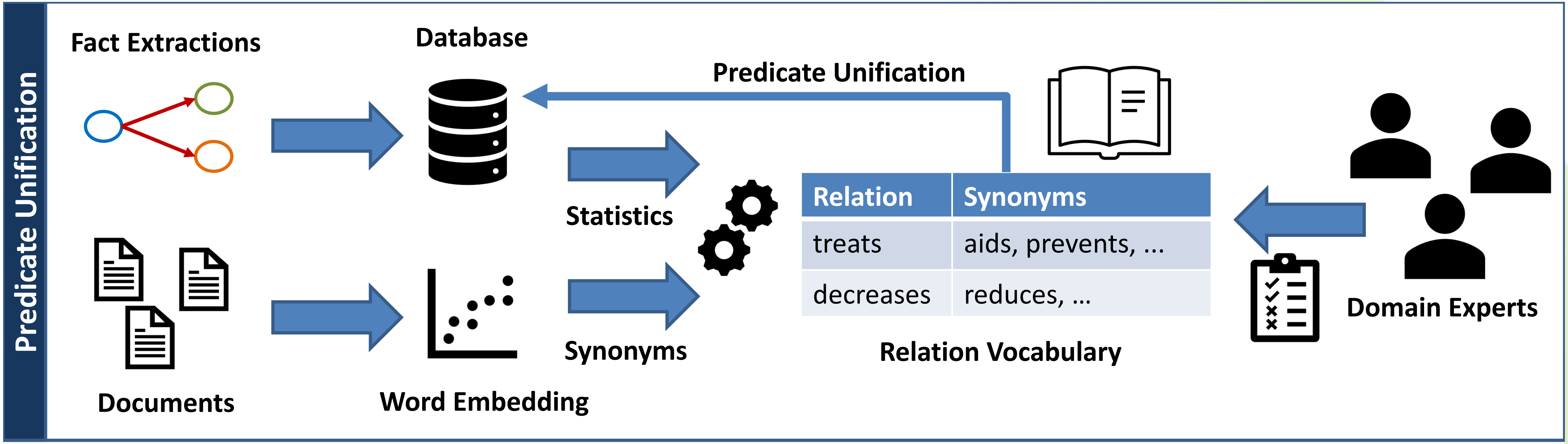
Unsupervised Information Extraction

- Extraction via **Open Information Extraction**:
  - **Stanford CoreNLP** (fast and reliable)
  - **OpenIE 6** (more precise but cost-intensive)
  - Toolbox can filter extractions by detected entities

- Extraction via **PathIE** (Path-based extraction method):
  - A statement is extracted if two entities are connected via a verb / keyword on the closest grammatical structure of a sentence
  - **PathIE** is recall-oriented and requires entity information

Predicate Unification & Constraints

- An iterative predicate unification algorithm maps synonymous predicates to precise relations



- Cleaning by **type integrity constraints**, e.g., $treats \subseteq Drug \times Disease$

Evaluation

- Our toolbox achieves a **moderate quality** in entity linking and information extraction
  - More details can be found in our paper

TABLE I
ENTITY LINKING QUALITY ON BIOMEDICAL BENCHMARKS: STATE-OF-THE-ART (SOTA) TAGGERS ARE COMPARED TO OUR UNSUPERVISED ENTITY LINKER. THE SOTA-TAGGING QUALITY RESULTS ARE FROM TAGGERONE [10] AND GNORMPLUS [9].

Benchmark	**Entity Type**	**Quality of SOTA Entity Linker**				**Quality of our Entity Linker**		
		Name	**Precision**	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
NCBI Disease [21]	Disease	TaggerOne	82.2 %	79.2 %	80.7 %	74.5 %	55.1 %	63.3 %
BioCreative V CD-R [22]	Disease	TaggerOne	84.6 %	82.7 %	83.7 %	82.8 %	62.0 %	70.9 %
BioCreative V CD-R [22]	Chemical	TaggerOne	88.8 %	90.3 %	89.5 %	76.6 %	78.7 %	77.6 %
BioCreative II GN [23]	Human Gene	GNormPlus	87.1 %	86.4 %	86.7 %	60.1 %	52.4 %	56.0 %

TABLE III
CDR2015 BENCHMARK EVALUATION [22]. THE TABLE REPORTS THE EXTRACTION QUALITY FOR OPENIE TOOLS, PATHIE AND BASELINES.

Method	**Quality**		
	Prec.	Rec.	F1
CoreNLP OpenIE	64.9 %	5.8 %	10.6 %
OpenIE6	53.1 %	5.5 %	10.0 %
PathIE	50.8 %	31.7 %	39.1 %
PathIE Stanza	51.1 %	30.9 %	38.5 %
Workshop Best Precision [22]	90.5 %	80.8 %	85.4 %
Workshop Best Recall [22]	86.1 %	86.2 %	86.1 %

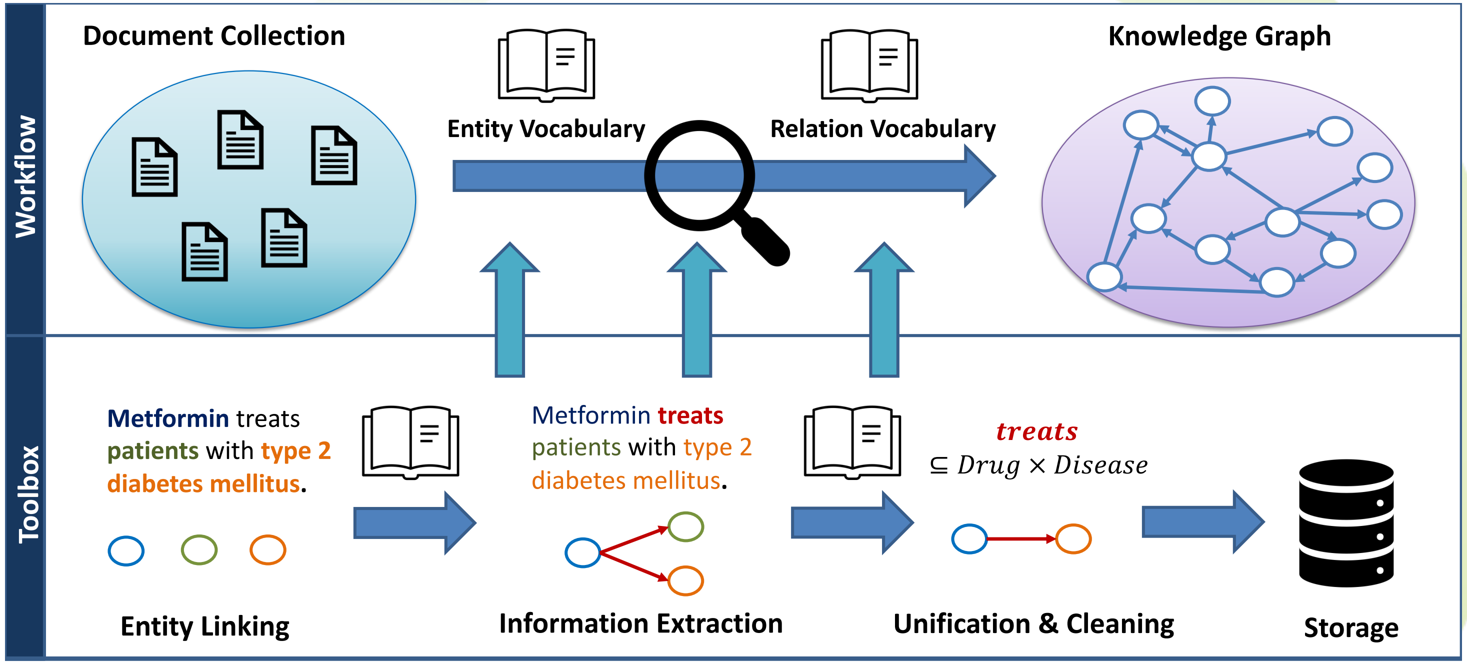
TABLE IV
BIOCREATIVE VI CHEMPROT EVALUATION [25]. THE TABLE REPORTS THE EXTRACTION QUALITY FOR OPENIE, PATHIE AND BASELINES.

Method	**Quality**		
	Prec.	Rec.	F1
CoreNLP OpenIE	59.3 %	5.1 %	9.3 %
OpenIE6	55.9 %	6.2 %	11.1 %
PathIE	30.3 %	55.3 %	39.1 %
PathIE Stanza	29.4 %	56.6 %	38.7 %
Sentence Co-Mention [25]	4.4 %	98.0 %	0.08 %
Workshop Best Precision [25]	74.4 %	55.3 %	63.4 %
Workshop Best Recall [25]	56.1 %	67.8 %	61.4 %
BioBERT [13]	77.0 %	75.9 %	76.5 %

Conclusion

- **Supervised** entity linking and information extraction **outperform** our **toolbox**
  - But they require a cost-intensive acquisition of domain-specific training data

- Our **toolbox bypasses** the need of **training data** but requires two vocabularies
  - Thus, the **toolbox** can **enable workflows** which are otherwise too cost-intensive to concern



Thank You!

FACHINFORMATIONSDIENST
PHARMAZIE
TU Braunschweig

Technische Universität Braunschweig

If you have any questions, contact me via:

kroll@ifis.cs.tu-bs.de

@HermannKroll