A Library Perspective on Nearly-Unsupervised Information Extraction Workflows in Digital Libraries

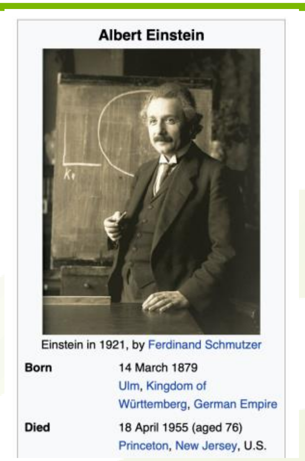
at JCDL2022

Hermann Kroll, Jan Pirklbauer,
Florian Plötzky and Wolf-Tilo Balke
Institut für Informationssysteme
Technische Universität Braunschweig

Motivation

ORKG

WIKIDATA



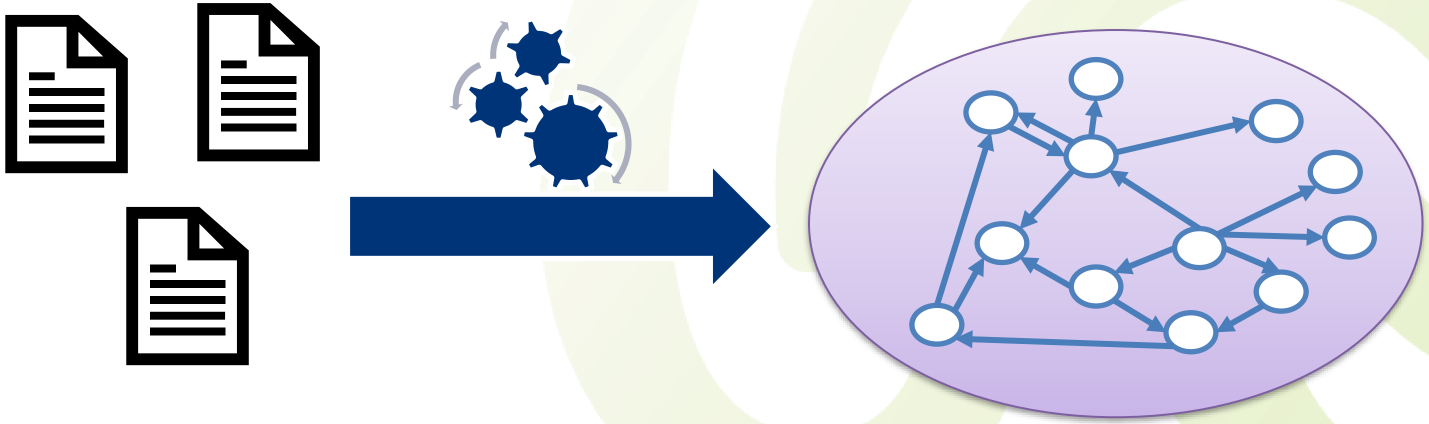
Albert Einstein

Einstein in 1921, by Ferdinand Schmutzer

Born	14 March 1879 Ulm, Kingdom of Württemberg, German Empire
Died	18 April 1955 (aged 76) Princeton, New Jersey, U.S.

→

DBpedia



Knowledge Graph

Nearly-Unsupervised Extraction Workflows

- *“The JCDL conference 2022 is held as a hybrid event in Cologne, Germany.”*

↓ Open Information Extraction

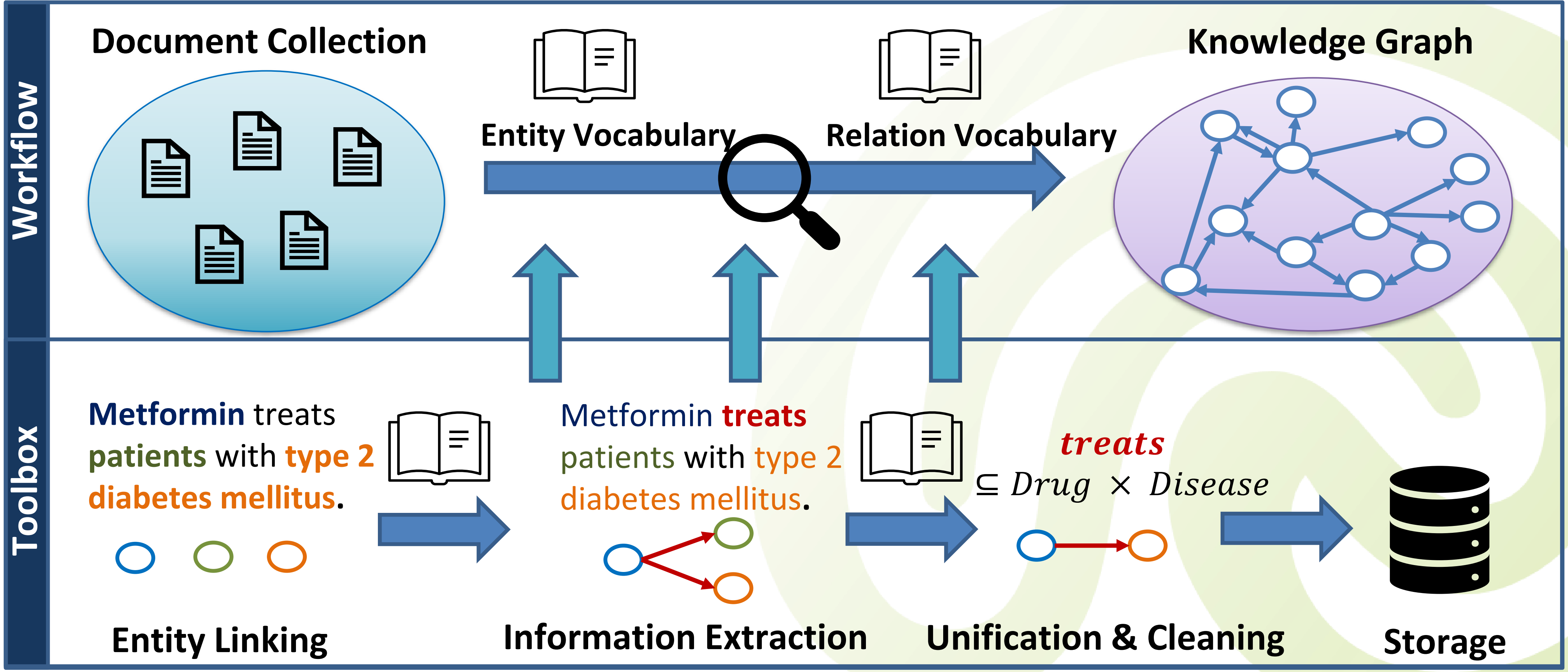
- *(The **JCDL conference 2022**; is held; as a hybrid event in **Cologne, Germany**)*

↓ Filtering

- *(**JCDL Conference 2022**; is held; **Cologne, Germany**)*

A Nearly-Unsupervised Extraction Toolbox

- Published at JCDL2021:
  - https://github.com/HermannKroll/KGExtractionToolbox
  - Shared as **Open Source**, written in **Python** and published with an **MIT license**



Research Questions

1. How much **expertise** and **effort** is required to apply nearly-unsupervised extractions across different domains?

2. How **generalizable** are these state-of-the-art extraction methods and particularly, how **useful** are the extraction results?

3. What is **missing** towards a **comprehensive information** extraction from texts, e.g., for retaining the original information?

Case Studies

- Investigated domains:
  - **Wikipedia** (descriptive writing, vocabularies available)
  - **Pharmacy** (entity-centric, controlled vocabularies)
  - **Political Sciences** (focused on topics and events, no vocabularies)

- Investigated methods:
  - Dictionary-based **entity linking** & Stanford Stanza **NER**
  - **PathIE** and **Open IE6** (2020)
  - **Filtering** (exact, partial, subject, no)
  - **Canonicalization** (vocabulary, word embedding)

Filtering Extractions

*(The **JCDL conference 2022**; is held; as a hybrid event in **Cologne, Germany**)*

- No Filter:

  (The JCDL conference 2022; is held; as a hybrid event in Cologne, Germany)
- Partial Filter:

  (JCDL conference 2022; is held; Cologne, Germany)
- Exact Filter:

  No Extraction
- Subject Filter (New):

  (JCDL conference 2022; is held; as a hybrid event in Cologne, Germany)

Summary Entity Linking

- Dictionary-based entity linking:
  - Derived vocabularies from Wikidata, MeSH, etc. were suitable
  - Short entity names were often linked incorrectly (homonyms)
  - Worked well in pharmacy (unambiguous concepts)

- Stanza NER:
  - Worked well for persons, organizations, countries, etc.
  - Did not produce precise entity identifiers
  - Struggled with bad metadata (e.g., abstracts in upper case)

Summary Information Extraction

- Open IE6:
  - Worked **well** for **short** but **bad** for **complex** sentences
  - Either noun phrases were short (good) or long (hard to filter)
  - Missed relations if they are not mentioned via a verb phrase, e.g., language from “*The German book Känguru-Chroniken*”

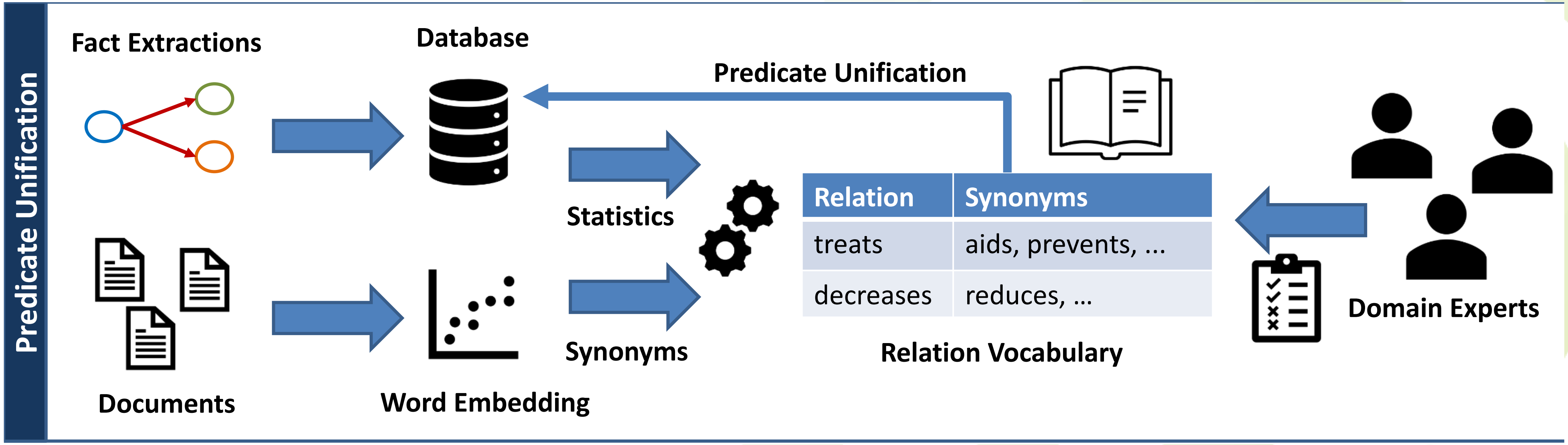
- PathIE:
  - Worked **well** if relations are **directed** (Person received Award) and **bad** if relations are **undirected** (Disease causes Disease)
  - Allowed extractions via special words (therapy, member of, …)

Summary Filtering

- **No** Filter:
  - No precise semantics
- **Partial** Filter:
  - Struggled for long noun phrases (complex sentences)
- **Exact** Filter:
  - Good quality but limited recall
- **Subject** Filter (New):
  - Allowed extraction of semi-structured information, e.g., actions performed by Albert Einstein or the EU

Summary Canonicalization

- **Building vocabularies** was challenging:
  - Worked well for: *treats, award received, member of, …*
  - Which relation is expressed by *do, publish, use, …?*
  - Sentence **context** was missing & embeddings did not help



Relation	Synonyms
treats	aids, prevents, …
decreases	reduces, …

Research Questions (1/3)

- How much **expertise** and **effort** is required to apply nearly-unsupervised extractions across different domains?

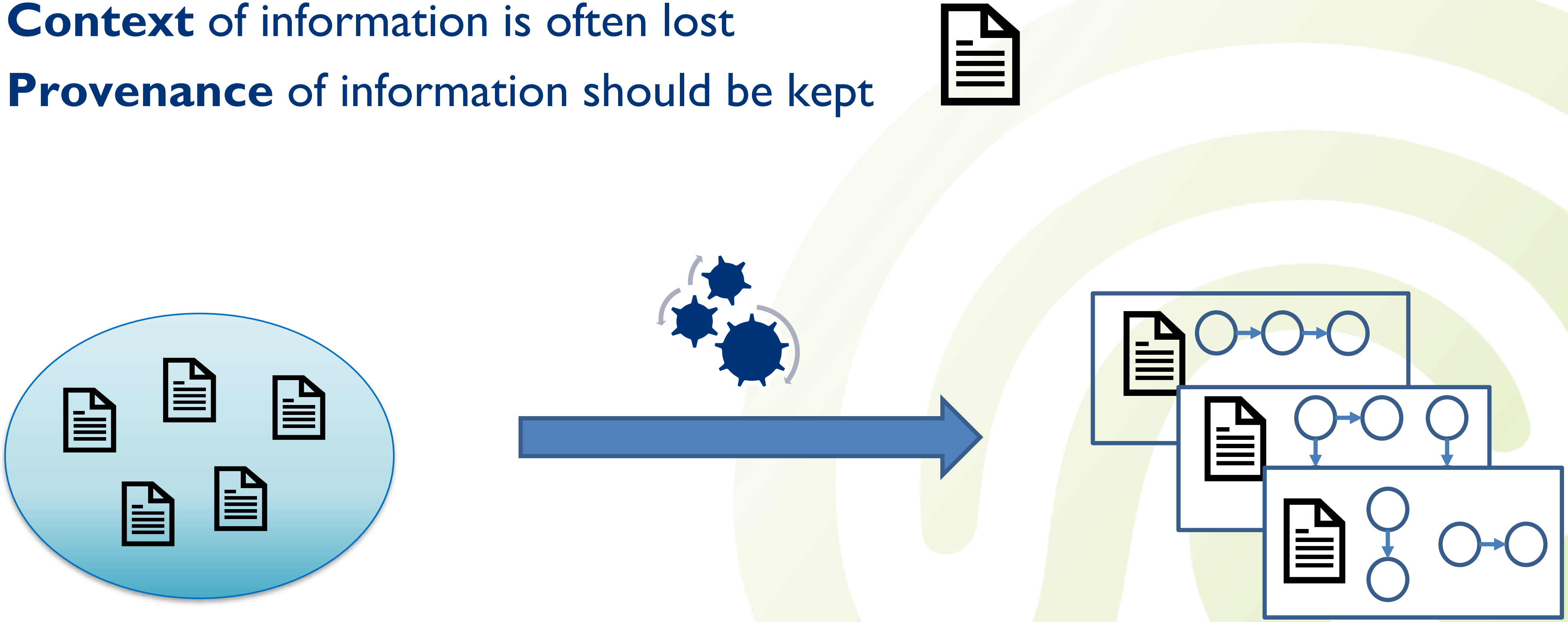
FACHINFORMATIONSDIENST PHARMAZIE TU Braunschweig	POLLUX	
www.narrative.pubpharm.de	Semi-structured knowledge	Some relations + Semi-structured knowledge
9x 2h sessions with experts Several weeks of development	4x 1.5h sessions with experts One week development	Three days

Research Questions (2/3)

- How **generalizable** are these state-of-the-art extraction methods and particularly, how **useful** are the extraction results?
  - Unsupervised extraction methods have a **moderate precision** but strongly **limited recall** (relations must be expressed via verbs)
  - **Filtering** is **necessary** to obtain **precise** relation semantics
  - **Entity** detection determines the **overall quality**
  - **Canonicalization** remains challenging and worked only in a few cases

Research Questions (3/3)

- What is **missing** towards a **comprehensive information** extraction from texts, e.g., for retaining the original information?
  - **Context** of information is often lost
  - **Provenance** of information should be kept



Best Practices

1. **Entity detection** is **required**

Metformin treats **patients** with **type 2 diabetes mellitus**.

2. Short and simple sentences are handled well and for long sentences use **exact** or **subject filter**

3. For relations that are **not** expressed via **verbs**, use **PathIE** + a relation vocabulary of special words

4. Use **PathIE** only if your relations are **directed**

$$\textit{treats} \subseteq \textit{Drug} \times \textit{Disease}$$

5. Otherwise, you will need supervision

Conclusion

- Nearly-Unsupervised workflows **are worth studying** in digital libraries because they
  - **Bypass training data** in the extraction phase completely
  - **Allow novel access** paths to digital libraries
  - **But require** extensive filtering in practice

FACHINFORMATIONSDIENST
PHARMAZIE
TU Braunschweig

www.narrative.pubpharm.de

Thank You!

FACHINFORMATIONSDIENST
PHARMAZIE
TU Braunschweig

Technische
Universität
Braunschweig

If you have any questions, contact me via:

✉ kroll@ifis.cs.tu-bs.de

@HermannKroll