Kroll, Hermann; Pirklbauer, Jan; Plötzky, Florian; Balke, Wolf-Tilo

A detailed library perspective on nearly unsupervised information extraction workflows in digital libraries

International Journal on Digital Libraries

Provided in Cooperation with:
Springer Nature

Suggested Citation: Kroll, Hermann; Pirklbauer, Jan; Plötzky, Florian; Balke, Wolf-Tilo (2023) : A detailed library perspective on nearly unsupervised information extraction workflows in digital libraries, International Journal on Digital Libraries, ISSN 1432-1300, Springer, Berlin, Heidelberg, Vol. 25, Iss. 2, pp. 401-425, https://doi.org/10.1007/s00799-023-00368-z

This Version is available at:
https://hdl.handle.net/10419/312295

International Journal on Digital Libraries (2024) 25:401–425
https://doi.org/10.1007/s00799-023-00368-z

A detailed library perspective on nearly unsupervised information extraction workflows in digital libraries

Hermann Kroll[1] · **Jan Pirklbauer**[1] · **Florian Plötzky**[1] · **Wolf-Tilo Balke**[1]

Received: 30 October 2022 / Revised: 9 May 2023 / Accepted: 19 May 2023 / Published online: 13 June 2023

Abstract
Information extraction can support novel and effective access paths for digital libraries. Nevertheless, designing reliable extraction workflows can be cost-intensive in practice. On the one hand, suitable extraction methods rely on domain-specific training data. On the other hand, unsupervised and open extraction methods usually produce not-canonicalized extraction results. This paper is an extension of our original work and tackles the question of how digital libraries can handle such extractions and whether their quality is sufficient in practice. We focus on unsupervised extraction workflows by analyzing them in case studies in the domains of encyclopedias (Wikipedia), Pharmacy, and Political Sciences. As an extension, we analyze the extractions in more detail, verify our findings on a second extraction method, discuss another canonicalizing method, and give an outlook on how non-English texts can be handled. Therefore, we report on opportunities and limitations. Finally, we discuss best practices for unsupervised extraction workflows.

Keywords Open information extraction · Extraction workflows · Digital libraries

1 Introduction

This paper is an extended version of our previous work [17] focusing on nearly unsupervised information extraction workflows in digital libraries. Extracting structured information from textual digital library collections enables novel access paths, e.g., answering complex queries over knowledge bases [2, 30], providing structured overviews about the latest literature [9], or discovering new knowledge [8].

However, utilizing information extraction (IE) tools in digital libraries is usually quite cost-intensive, which hampers the implementation in practice. On the one hand, extraction methods usually rely on supervision, i.e., ten thousands of examples must be given for training suitable extraction models [35]. On the other hand, utilizing the latest natural language processing (NLP) tools in productive pipelines requires high expertise and computational resources.

In addition to supervised IE, Open IE methods (OpenIE) have been developed to work out-of-the-box without additional domain-specific training [11, 23]. But why aren't they used broadly in digital library applications? The reason is that OpenIE generates non-canonicalized (not normalized) results, i.e., several extractions describing the same piece of information may be structured in completely different ways (synonymous relations, paraphrased information, etc.). But such non-canonicalized results are generally not helpful in practice, because a clear relation and entity semantics like in supervised extraction workflows is vital for information management and query processing. Since the lack of clear semantics has been recognized as a major issue, cleaning and canonicalization methods have been investigated to better handle such extractions [31]. Still are they ready for application in digital libraries?

In this paper, case studies are used to find out how suitable nearly unsupervised methods are to design reliable extraction workflows. In particular, we analyze extraction and cleaning methods from the perspective of a digital library

✉ Hermann Kroll
kroll@ifis.cs.tu-bs.de

✉ Jan Pirklbauer
j.pirklbauer@tu-bs.de

✉ Florian Plötzky
ploetzky@ifis.cs.tu-bs.de

✉ Wolf-Tilo Balke
balke@ifis.cs.tu-bs.de

[1] Institute for Information Systems, TU Braunschweig, Mühlenpfordtstr. 23, Braunschweig 38106, Lower Saxony, Germany

by assessing the required expertise, domain knowledge, computational costs and result quality.

Therefore, we selected our toolbox for a nearly unsupervised extraction from text published in JCDL 2021 [15]. The toolbox contains interfaces to the latest named entity recognition (NER) and open information extraction methods. In addition, it includes cleaning and canonicalization methods to handle noisy extractions by utilizing domain-specific information. Our corresponding paper [15] advertises the toolbox to considerably decrease the need for supervision and to be transferable across domains; nevertheless, it comes with several limitations:

1. Although we did report on the extraction quality (good precision, low recall), we did **not** report on the **costs of applying the toolbox**, i.e., how much expertise and computational costs are required for a reliable workflow.
2. We applied the toolbox only in the biomedical domain, which lessens the **generalizability of our findings**.
3. Moreover, we did **not** report what is **technically and conceptually missing** in such extraction workflows.
4. We focused on **English** texts and did not analyze workflows for **non-English** texts yet.

In this paper, we address the previous issues by analyzing the toolbox application in three distinct real-world settings from a library perspective: 1. We extracted knowledge about scientists from the online encyclopedia Wikipedia (controlled vocabularies, descriptive writing). 2. We applied the toolbox to the pharmaceutical domain (controlled vocabularies, entity-centric knowledge) in cooperation with the specialized information service for Pharmacy (www.pubpharm.de). 3. We applied the toolbox in Political Sciences (open vocabulary, topic/event-centric knowledge) in cooperation with the specialized information service for Political Sciences [29] (www.pollux-fid.de). For Pharmacy and Political Sciences, we recruited associated domain experts for expertise in the evaluation. We performed these three case studies to answer the following 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 toward a comprehensive information extraction from texts, e.g., for retaining the original information?*

In addition to those questions, we discuss how digital libraries may handle non-English texts with our toolbox. This paper is an extended version of our previous article [17]: For our extension, we (1) give more insights and details for each case study in Sect. 4, (2) investigate the complexity of extracted noun phrases in Sect. 4.4, (3) apply and analyze a second OpenIE tool, namely CoreNLP OpenIE, to generalize our findings in Sect. 4.5, (4) have a close look on an unsupervised canonicalization method for verb phrases in Sect. 5, and (5) dive into machine translation to apply the toolbox on non-English texts, at the example of German in Sect. 6. For a comparison of old and new hardware, we also measured the runtimes on our latest server from 2021 in Sect. 7.5.1.

2 Related work

The main goal of information extraction (IE) is the extraction of structured information from unstructured or semi-structured information such as texts, tables, figures, and more [11, 22, 23, 35]. In the following, we give an overview of challenges and research trends in IE from texts.

Current Trends. Modern IE research mainly focuses on improving the extraction accuracy, which is typically measured on benchmarks [3, 11]. Indeed, previous evaluations have shown that IE methods already produce good results, but the research is still ongoing [3, 5, 11, 15, 26]. Primarily driven by the development of language models like BERT [5], IE has made a step forward.

However, these systems rely on supervised learning and thus need large-scale training data that cannot be reliably transferred across domains. In brief, although supervised methods are up to the job with reasonable quality, their practical application comes at high costs. The expenses for supervision lead to the design of zero-shot, semi-supervised, and distant supervised extraction methods (see [35] for a good overview).

Open Information Extraction. Instead of designing extraction systems for each domain, methods like unsupervised information extraction (OpenIE) are proposed to change the game [26]. OpenIE aims to extract knowledge from texts without knowing the entity and relation domains a-priori [26, 35]. While supervised (closed) methods focus on domain-specific and relevant relations and concepts, open methods are more flexible and may be applied across domains [26, 35].

Canonicalization of OpenIE. Vashishth proposed CESI to canonicalize OpenIE extractions by clustering noun and verb phases with the help of side information [31]. However, CESI was analyzed for short phrases that refer to precise entities. In addition, studies have shown that OpenIE methods may struggle to handle scientific texts well because sentences are often long and domain-specific vocabulary terms are used [7]. While research in both directions (open and closed) is still ongoing, some works bridge the gap between both worlds: Kruiper et al. propose the task of Semi-Open Relation extraction [20], i.e., they use domain-specific information

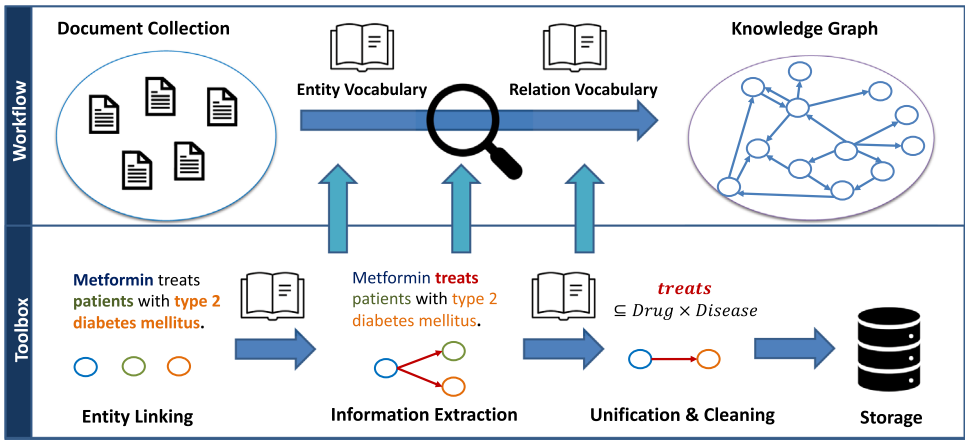

Fig. 1 The Toolbox's Systematic Overview: Entity linking detects concepts/entities, and information extraction extracts relations between them. Then, the output will be cleaned and loaded into a structured repository [15]

to filter irrelevant open information extractions. Similarly, we showed that domain-specific filtering of OpenIE outputs could yield helpful results [15].

Information Extraction in Digital Libraries. Digital libraries are interested in practical IE workflows to allow novel applications; see this tutorial at JCDL2016 [36]. IE can allow literature-based discovery workflows, which have been studied on DBpedia [30]. The extraction of entities and relations is therefore challenging. That is why modern approaches build upon language models and supervision for a reliable extraction [28]. These language models require extensive computational resources for training and application [5, 21]. Good examples for IE are DBpedia [2], which was harvested from Wikipedia infoboxes or the SemMedDB, which is a collection of biomedical statements harvested from PubMed [10, 37]. Hristovski et al. have used the SemMedDB to perform knowledge discovery [8]. Nevertheless, the construction of SemMedDB required biomedical experiences to define hand-written rules for the extraction. In contrast to the previous works, our work focused on nearly unsupervised extraction workflows that do not rely on training data for the extraction phase.

3 Study objectives

In the following, we briefly summarize the nearly unsupervised extraction toolbox, raise research questions for our case studies, and explain why we selected the three domains here. A systematic toolbox overview is shown in Fig. 1. Our main objective here is to analyze unsupervised extraction workflows from a digital library perspective.

3.1 Overview of the toolbox

The extraction toolbox covers three common IE areas: entity detection, information extraction, and canonicalization. We shared our toolbox as open-source software and made it publicly available.[1,2] We focus on this toolbox because it proposed an eased and nearly unsupervised extraction workflow by integrating the latest unsupervised extraction plus suitable cleaning methods.

Nearly Unsupervised. We call an information extraction workflow nearly unsupervised if two conditions hold: 1. No training data are required to train or fine-tune an entity detection or information extraction model. In other words, entities and statements are extracted without supervision. And 2. entity information and a relation vocabulary are used to clean not-normalized extraction outputs, e.g., by filtering OpenIE noun phrases via detected entities or canonicalizing synonymous verb phrases to precise relations. In contrast to pure unsupervised workflows, our workflow requires the design of an entity and relation vocabulary to obtain precise relation semantics, e.g., a treats relation between drugs and diseases.

Entity Detection. The toolbox integrates interfaces to one of the latest NER tools, Stanford Stanza [27]. Stanford Stanza is a pre-trained neural model that can be applied without adapting it to a certain domain. Stanza is capable of detecting 18 general-purpose entity types like *persons*, *organizations*, *countries*, and *dates* in texts; see [27] for a complete overview. In addition, the toolbox supports the linking of custom entity vocabularies via a dictionary-based lookup method. The entity linker supports an abbreviation resolution and handling of short homonymous terms (link if the entity is mentioned with a longer mention in the text).

Information Extraction. The toolbox integrates implements interfaces to OpenIE methods, Stanford CoreNLP [23] and OpenIE6 [11]. Besides, the toolbox includes a self-developed path-based extraction method named PathIE. PathIE extracts statements between entities in a sentence if connected in the grammatical structure via verb phrases or custom keywords (e.g., treatment, inhibition, award, and

[1] https://github.com/HermannKroll/KGExtractionToolbox.

[2] https://archive.softwareheritage.org/swh:1:dir:5b575ac043e2bd61999250564a16a220c88ee5c9.

member of) that can be specified beforehand. The OpenIE methods work entirely without entity information, whereas the PathIE requires entity annotations as starting points (as an input).

Cleaning and Canonicalization. OpenIE and PathIE may produce non-helpful and non-canonicalized (not-normalized) outputs, i.e., synonymous noun and verb phrases that describe the same information. The toolbox supports canonicalizing and filtering such outputs automatically. First, extracted noun phrases can be filtered by entity annotations, i.e., only noun phrases that include relevant entities are kept. Here, three different filters are supported to filter noun phrases: exact (noun phrase matches an entity), partial (noun phrase partially includes an entity), and no filter (keep original noun phrase). We will introduce the subject filter as a new option in our case studies. For convenience, the subject filter requires the extracted subject noun phrase to be a detected entity. And it keeps the object noun phrase as it is. As a recent example, consider the sentence: ***Queen of England*** passed away in **2022** after a long reign in **Balmoral Castle**. Assume that we detected the bold text spans as entities. For the following extraction (Queen of England; passed away; in 2022 after a long reign in Balmoral Castle), filtering will then yield:

No Filter:	Keep the extraction as it is.
Partial Filter:	(Queen of England; passed away; Balmoral Castle) and (Queen of England; passed away; 2022).
Exact Filter:	will not return anything because the object consists of more than the detected entity.
Subject Filter:	(Queen of England; passed away; after a long reign in Balmoral Castle)

Second, an iterative cleaning algorithm is integrated that can canonicalize synonymous verb phrases to precise relations, e.g., *birthplace* or *place of birth* to *born in*. Therefore, users can export statistics about the non-canonicalized verb phrases and build a so-called relation vocabulary. Each entry of this vocabulary is a relation consisting of a name and a set of synonyms. The toolbox utilizes this vocabulary to automatically map synonymous verb phrases to precise relations. Word embeddings are supported in the canonicalization procedure to bypass an exhausting editing of the relation vocabulary. The central idea of word embeddings is that words with a similar context appear close in the vector space [25]. The word embedding is then used to automatically map a new verb phrase to the closest match (most similar) in the vocabulary. Relation type constraints can then be used to filter the extractions further, i.e., a relation type constraint describes which entity types are allowed as subjects and objects. For example, *born in* can be defined as a relation between *persons* and *countries*. Other extractions that hurt these constraints are then removed. We already reported on some challenges of OpenIE extractions, especially on handling noun phrases [14]. In contrast to our previous works, this work analyzes the complete workflow in three domains from a library perspective.

3.2 Study goals

The study goals concern three concrete areas of study: 1. application costs, 2. generalizability, and 3. limitations for a comprehensive IE. However, answering these questions on a purely quantitative level is challenging, e.g., how can the costs be measured? That is why we report our findings as a mixture of quantitative measures (e.g., time spent and runtimes) and qualitative observations (what works well and what does not). We define evaluation criteria for all of the three aspects in the following.

3.2.1 Application costs

We understand everything necessary to implement a workflow with the toolbox as *application costs*. We estimate the application costs in terms of

Data Preparation:	transforming data into toolbox formats (e.g., JSON), working with toolbox outputs (TSV/JSON)
Implementation:	computational costs (runtime and space), scalability, executed steps, effort to choose parameters, encountered issues
Domain Knowledge:	entity and relation vocabulary design, required knowledge for canonicalization

3.2.2 Generalizability

In short, how well are the proposed methods generalizable across domains, and how useful are the results?

Extraction quality:	benchmarks (precision and recall), observations, extraction limitations
Usefulness:	relevance of statements (e.g., non-obvious statements), domain insights, helpfulness for domain experts, usefulness in applications

Information, originally connected in coherent written texts, might be broken into not helpful pieces in the end. For a good example, consider a drug-disease treatment: Here context information like the dose or treatment duration, which could give more information about the statement's validity [13], might get lost. We refer to such information as the **context** of statements, e.g., the surrounding scope in which a

Table 1 The number of documents and sentences is reported for each collection and sample

Collection	Size	Sample	
		#Documents	*#Sentences*
English wikipedia	6.3 M	2373	74.5k
PubMed	33 M	10k	87.1k
Political sciences	1.7 M	10k	66.9k

statement is valid. We already discussed why context information is essential when extracting statements; see [13, 18]. In addition, the connection between statements might get lost, too, e.g., an assumption might lead to a conclusion. We call this the **coherence of statements**. They are crucial for real-world applications, but have they being considered yet?

3.2.3 On context and coherence

Contexts affect the validity of statements, and coherence describes how statements belong together. We evaluate the following criteria:

Contexts: relevance of contexts, which kind of information requires context, how does the context affect the validity of extracted statements, what must be done to retain context

Coherence: complex information that is broken into pieces, which kind of information is broken down, what are the subsequent problems with such a decomposition

3.3 Case study selection

We applied the toolbox in three different domains to generalize the findings in this paper. Here we focused on natural language texts written in the English language. We describe the domains and their characteristics in the following. Table 1 provides statistics about the used data and samples.

3.3.1 Wikipedia

A prime example of an encyclopedia is the free and collaborative Wikipedia. Encyclopedic texts should be written in descriptive and objective language, i.e., wording and framing should not play any role. Wikipedia captures knowledge about certain items (persons, locations, events, etc.), in our understanding, entities. Here controlled ontologies about entities and relations are available; see Wikidata [32] as a good example. However, Wikipedia texts also tend to include very long and complex sentences. For this case study, we focus on knowledge about famous fictional and non-fictional scientists (about 2.4k scientists with an English Wikipedia article and Wikidata entry). This case study was selected because sentences are written objectively, and controlled vocabularies are available for usage.

3.3.2 Pharmaceutical domain

The pharmaceutical domain focuses on entity-centric knowledge, i.e., statements about entities such as drugs, diseases, treatments, and side effects. Many vocabularies and ontologies are curated to describe relevant biomedical entities, e.g., the National Library of Medicine (NLM) maintains the so-called Medical Subject Headings (MeSH).[3] These headings are entities with descriptions, ontological relations (subclasses), and suitable synonyms. In this paper, we select a subset of the most comprehensive biomedical collection, the NLM Medline collection.[4] Medline includes around 35 million publications with metadata (title, abstracts, keywords, authors, publication information, etc.). The specialized information service for Pharmacy was interested in statements about drugs. Therefore, we applied the entity linking step to all Medline abstracts (Dec. 2021) and randomly picked a subset of 10k abstracts that included at least one drug mention.

3.3.3 Political sciences

The Political Sciences domain encompasses a diverse range of content, e.g., publications about topics and events, debates, news, and political analyses. Because of its diversity, this domain does not provide extensive curated vocabularies and ontologies. We argue that entity subsets of knowledge bases like Wikidata [32] or DBpedia [2] might be good starting points to derive some entity vocabularies regarding persons, events, locations, and more. Still, Wikidata and DBpedia are built as general-purpose knowledge bases. They are thus not focused on Political Sciences (in contrast to MeSH for the biomedical domain). Nevertheless, they might be helpful to analyze texts in Political Sciences, which is why we analyze them for a practical application here. In addition, descriptions of entities in Political Sciences tend to be subjective, i.e., they depend on different viewpoints and schools of thought. For example, the accession of Crimea to Russia in 2014 was a highly discussed topic, whether this event could be seen as peaceful secession or as an annexation. In contrast to objective and entity-centric statements in biomedicine, Political Sciences are far more based on the wording and framing of certain events. This case study analyzes how far IE methods can bring structure into these texts and where these methods fail. The specialized information service for Political

[3] https://meshb.nlm.nih.gov/search.

[4] https://www.nlm.nih.gov/medline/medline_overview.html.

Table 2 Corpus and entity detection statistics for our case studies

	Sentences		Entity detection	
	#Sent.	*#with2E*	*#NER*	*#EL*
Wikipedia	74.5k	50.3k	155.0k	113.2k
Pharmacy	87.1k	47.4k	–	232.5k
Pol. Sci.	66.9k	17.6k	80.0k	3.7k

We report the number of sentences, sentences with at least two entities mentions, Stanza NER, and entity linking annotations

Sciences (Pollux) provided us with around three million publications (around 1.3 million English abstracts). Our case study is based on a random sample of 10k abstracts selected from the English subset. In addition, domain experts manually selected five abstracts due to their focus on the diverse topics of the EU, philosophy, international relations, and parliamentarism (Tables 2 and 3).

4 Case studies

For our case studies, we developed scripts, produced intermediate results, and implemented some improvements to the toolbox. The details, used data, and produced results of every case study can be found in our evaluation scripts on GitHub (see the Toolbox GitHub Repository). We included a Readme file[5] to document the following case studies. All our experiments and time measurements were performed on our server, having two Intel Xeon E5-2687W (@3,1GHz, eight cores, 16 threads), 377GB of DDR3 main memory, one Nvidia 1080 TI GTX GPU, and SSDs as storage.

For the first part of this section, we used OpenIE6 to perform the OpenIE extractions because it was the latest OpenIE tool available in the toolbox. To better generalize those findings, we subsequently analyzed the produced noun phrases in detail and compare the results to the CoreNLP OpenIE tool; see Sect. 4.4 and Sect. 4.5.

4.1 Wikipedia case study

This first case study was based on 2.3k English Wikipedia full-text articles about scientists. The conversion of Wikipedia articles was simple: We downloaded the available English Wikipedia dump (Dec. 2021), used the WikiExtractor [1] to retrieve plain texts, and filtered these texts by our scientist's criteria (title must be about a scientist of Wikidata). Next, we developed a Python script to transform the plain texts into a JSON format for the toolbox. The data transformations took half a person-day.

[5] https://github.com/HermannKroll/KGExtractionToolbox/blob/main/README_CASE_STUDIES.md.

4.1.1 Entity linking

In this case study, we focused on statements about scientists, such as works, scientific organizations, and degrees. Therefore, we performed entity linking to identify these concepts and use them to filter the extraction outputs. We derived corresponding entity vocabularies from Wikidata by utilizing the official SPARQL endpoint. We retrieved vocabularies by asking for English labels and alternative labels for the following entity types: *Academia of Sciences*, *Awards*, *Countries*, *Doctoral Degrees*, *Religions* and *Irreligions*, *Scientists*, *Professional Societies*, *Scientific Societies* and *Universities*.

This query returned rows including the entity id, the entity name, and a;-separated list of English alternative labels for the corresponding entity. We adjusted the SPARQL queries to directly download the vocabularies as TSV files in the toolbox format. A first look over this entity vocabulary revealed some misleading labels (e.g., the, he, she, and, or), which we removed. Our final vocabulary included 27,864 distinct entities and 68,668 distinct terms.

We applied the dictionary-based entity linker utilizing our vocabulary to the articles. The linker yielded many erroneously linked entities because of very ambiguous labels in the dictionary, e.g., the mentions *doctor*, *atom*, and *observation* were linked to fictional characters which are scientists regarding the Wikidata ontology. Next, synonyms like *Einstein* were erroneously linked when talking about his family or talking about the term *Einstein* in the sense of *genius*. The linker also ignored pronouns completely, i.e., no co-reference resolution was applied. Especially in Wikipedia articles, pronouns are often used. In addition, we executed the NER tool Stanford Stanza to recognize general-purpose entity types like dates or organizations. A closer look at Stanza's results revealed that short entity names were too ambiguous. That is why we removed all detected entities with less than five characters. This step yielded 155k Stanza NER mentions and 113.2k dictionary-based entity links.

4.1.2 Information extraction

OpenIE6. We applied the OpenIE6 method and the entity filter methods (no filter, partial, exact). We obtained 117.1k (no filter), 317.8k (partial), and 2.9k (exact) extractions. Note that statements can be duplicated for the partial filter if multiple entities are included within the same noun phrase. We exported 100 results for each filter randomly and analyzed them. In the following, we report on some examples of good and bad extractions.

Some interesting results about Albert Einstein are listed in Table 4. OpenIE6 produced correct and helpful extractions when sentences were short and simple (no nested structure, no relative clauses, etc.). When sentences became longer, the tool yielded short subjects but long and complex objects,

Table 3 OpenIE6 extraction and filtering statistics: We report the percentage of complex subjects and objects, the number of extractions computed by the different entity filters (no, partial, exact, subject), and PathIE (number of extractions)

	OpenIE6						PathIE
	C. Subjs. (%)	*C. Objs.* (%)	*#No EF*	*#Part. EF*	*#Exact EF*	*#Subj. EF*	*#Extr.*
Wikipedia	16.2	74.5	177.1k	317.8k	2.9k	80.9 k	1.3 M
Pharmacy	37.8	72.1	207.6k	88.0k	291	15.1k[6]	430.8k
Pol. Sci.	32.0	74.3	147.2k	28.6k	128	7.3k	–

Table 4 OpenIE6 example extractions from the Wikipedia article of Albert Einstein. On the left, the corresponding entity filter is shown (subject, partial and exact). Subject[S], predicate[P] and object[O] are highlighted respectively

Wikipedia	Exact	E1.1	In 1933, while **Einstein**[S] **(Person)** *was visiting*[P] **the United States**[O] **(Country)**, [...]
		E1.2	On 30 April 1905, Einstein completed his thesis, with **Alfred Kleiner**[S] **(Person)**, *[be] Professor*[P] of **Experimental Physics**[O] **(ORG)**, serving as "pro-forma" advisor.
	Partial	E2.1	In a German-language letter to **philosopher**[O] **(Profession)** Eric Gutkind, dated 3 January 1954, **Einstein**[S] **(Person)** *wrote*[P]: [...]
		E2.2	**Einstein**[S] **(Person)** *was elected*[P] a Foreign Member of **the Royal Society**[O] **(Org)** (ForMemRS) in 1921.
	Subject	E3.1	During an address to Caltech's students, **Einstein**[S] **(Person)** *noted*[P] **that science was often inclined to do more harm than good**[O].
		E3.2	**Einstein**[S] **(Person)** *started teaching*[P] himself **calculus at 12**[O], and as a 14-year-old [...]

e.g., a whole subordinate clause like *that science was often inclined to do more harm than good*. See E3.1 in Table 4.

We developed a short script to quantify them to understand better how many subjects and objects were complex. Therefore, we formulated regular expressions to check if a sentence or noun phrase contained multiple clauses split by punctuation (,;:), or words (and, or, that, thus, hence, because, due, etc.). We then counted subjects and objects as complex if they matched one of these regular expressions. In addition, noun phrases that consumed more than 50% of the sentence were considered complex. And if noun phrases consumed more than 20% of the sentence and the sentence itself consisted of multiple clauses (regular expressions again), we denoted the noun phrases as complex. Note that we are aware of the limitations of such a heuristic. That is why we compared this heuristic to other methods in depth in Sect. 4.4. Returning to our sample, 16.2% of subjects and 74.5% of objects were classified as complex. We iterated over these classifications to verify the filter criteria.

Partial Entity Filter. This filter yielded problematic results because much information was lost, e.g., a whole subordinate clause was broken down into a single entity regardless of where the entity appeared in this clause. In some cases, this filtering completely altered the sentence's original information; see E2.2 for a good example. Here the extraction *Einstein was elected the Royal Society* was nonsense because *Foreign Member* was filtered out. In E2.1, the extracted statement missed that the *philosopher* was *Eric Gutkind*, and thus lost relevant information.

Exact Entity Filter. The exact filter was very restrictive because the number of extractions was reduced from 117.9 to 2.9k. However, the extraction seemed to have good quality. In E1.1, the extraction *Einstein was visiting the US* was correct, but the context about the year *1933* was lost. Extraction E1.2 showed that OpenIE6 was capable of extracting implicit statements like *be Professor of*. Again, the surrounding context about the year and Einstein was lost. Other extractions showed that a co-reference resolution would be beneficial to resolve mentions like *his*, *in the same article*, and *these models*.

Subject Entity Filter. We observed many complex object phrases (74.5% in sum). These complex phrases contained more information than a single entity. Filtering them led to many wrongly extracted statements. In contrast, subject phrases were often simple and might stand for a single entity (only 16.2% are complex). Because of these observations, we developed a subject entity filter, i.e., only subjects had to match entities directly. The idea was to identify subjects as precise entities and keep object phrases in their original form to retain all information.

Results. This filter worked as expected: In E3.1 and E3.2, the subject was identified as the Person *Einstein*, whereas the original information was kept in the object phrase. For example, this filtering allowed us to generate a structured overview of Albert Einstein: (excelled, at math from a young age), (published, hundreds of articles throughout his life), and

[6] We wrongly reported 151k in [17].

(attempted, to generalize his theory of gravitation following his research on general relativity).

PathIE. In addition to analyzing OpenIE6, we investigated how useful PathIE is in extracting relations between the relevant entity types, such as scientists and awards. PathIE allowed us to specify keywords that can indicate a relation. In a first attempt, we applied PathIE with a small relation vocabulary of Wikidata. We exported the English labels and alternative labels of eleven Wikidata properties that describe the relations between the given entity types: academic degree, award received, date of birth, date of death, field of work, member of, native language, occupation, religion, and writing language. For example, the entry *award received* had the following synonyms: award received, award won, awarded, awards received, honorary title, honors, honours, medals, prize awarded, prize received, recognition title, win, winner of, award, and awards.

We exported and evaluated 100 randomly selected PathIE extractions. When several entities were detected in long and nested sentences, PathIE yielded many wrong extractions because the corresponding entities were connected via some verb phrases, e.g., *Einstein return Zurich* from *Einstein visited relatives in Germany while Maric returned to Zurich* or *Written languages write Leningrad*. Filtering them by entity types like (Person, Date) or (Person, Award) revealed more helpful extractions, e.g., *Einstein win Nobel Prize* from *Einstein received news that he had won the Nobel Prize in November*.

However, we encountered severe entity linking issues when analyzing the cleaned OpenIE6 and PathIE extractions. On the one hand, ambiguous terms were linked wrongly. On the other hand, fragments of a text span were linked against an entity although the whole text span referred to a single entity, e.g., only linking *Albert Einstein* in the text mention *Albert Einstein's Theory of Relativity was published in 1916*. These issues directly affected the extraction quality. We stopped the extraction part at this point.

4.1.3 Canonicalization

We used our small relation vocabulary to canonicalize the extractions. This procedure did work out for PathIE because it directly extracted the vocabulary entries from the texts. For example, we could retrieve a list of statements that indicate an *award received* relation. However, further cleaning was required to obtain *award received* relations between persons and awards. We analyzed 100 entries for this relation. Although some extraction were correct, 60 of 100 extractions had linked awards that were not helpful, e.g., *awards*, *doctor*, *medal*, *president* and *master*. The remaining 40 extractions displayed six wrongly identified persons. However, the remaining 34 extractions seemed plausible, although some information was missed, like the *Nobel prize's* category.

Next, we used the same relation vocabulary to canonicalize the OpenIE6 extractions. In brief, the canonicalization procedure did not work. The reason was that the extracted verb phrases did not appear directly in the vocabulary, e.g., see the aforementioned terms for *award received*. Thus, we used a pre-trained English Wikipedia word embedding from fasttext[7] to find similar matches in the relation vocabulary. We adjusted the cleaning parameters (how similar terms must be and how often terms must occur) and canonicalized the OpenIE6 verb phrases. However, most verb phrases were mapped wrongly because the vocabulary was relatively small, e.g., *divorce* was mapped to *date of death* because it was the closest match (in terms of vector space similarity).

We then derived a list of 120 Wikidata properties that involved persons (ignoring usernames and identifiers) to find more matches. We repeated the canonicalization and analyzed 100 extractions obtained by the subject entity filter because it retrieved the most helpful results in the previous step. Most of the canonicalized verb phrases were mapped incorrectly, e.g., mapping *start teach* to *educated at* or *begin* to *death of place* was wrong. For a positive example, the verb phrase *publish* was mapped to the relation *notable work* and *write* to *author*, e.g., *Galileo publish* (↦ *notable work*) *Dialogue Concerning the Two Chief World Systems*. Although this relation was correct for a few extractions, most of these mappings were problematic, e.g., *Einstein publish* (↦ *notable work*) *his own articles describing the model among them.* Here the object phrase did not contain a notable work in the sense of how we would understand it.

In summary, the canonicalization procedure had many problems for OpenIE6 extractions. The main issue was that the canonicalization procedure only considered the verb phrase, not the surrounding context in a sentence. But this surrounding context is essential to determine the relation, e.g., the verb phrase *use* could refer to many different relations depending on a concrete sentence. In addition, the relation vocabulary obtained from Wikidata might be insufficient because it did not contain verb phrases as we would expect them. Wikidata describes relations by using substantives and nouns, e.g., notable work of, notable work by, notably created by for the relation *notable work*. However, such substantives should typically not be included in the verb phrase of an OpenIE extraction because they are not verbs.

4.1.4 Application costs

We spent much of our time understanding the Wikidata ontology and formulating suitable SPARQL queries to retrieve the utilized vocabularies. The corresponding vocabularies could

[7] https://fasttext.cc/docs/en/pretrained-vectors.html.

be exported directly from Wikidata and did not need transformations besides a concatenation of files. We formulated several SQL queries to analyze, clean, and filter entity annotations and extractions in the toolbox's underlying database. In summary, three persons performed this case study within three person-days.

4.1.5 Generalizability

We had a close look at existing Wikipedia relation extraction benchmarks for evaluation. Unfortunately, these benchmarks are often built distantly supervised, i.e., if two entities appear in a sentence, and both entities have a relation in a knowledge base, then this relation is the class that must be predicted for this sentence. In other words, the relation does not have to appear within the sentence. Furthermore, these benchmarks often require domain knowledge, e.g., if a football player started his career at a sports team, then the football player *played for* this team. This additional knowledge is typically not included in OpenIE methods. OpenIE extracts statements based on grammatical patterns in a sentence: For the previous example, the tool would extract that the football player started his career on the sports team but not that he also played for the team. So we did not evaluate the extraction tool on existing benchmarks because we had reason to expect the quality to be low by design. Moreover, mapping verb phrases to precise relations would also be too challenging. In contrast, we wanted to understand how useful the results were for practical applications.

First, an improved entity linking would have solved several issues in our case study. Next, the handling of complex noun phrases was an issue: Although the exact entity filter was too restrictive, it resulted in suitable extractions. The partial entity filter messed up the original information and was thus not helpful. OpenIE6 and the subject entity filter allowed us to retrieve a list of actions performed by Albert Einstein, for example. However, this filtering did not yield a canonicalized knowledge base by design. Our case study has shown that PathIE could extract relations between scientists and awards. Although we could not evaluate the quality in rough numbers, we spent three person-days designing a possible extraction workflow. Here, the toolbox allowed us to retrieve such semi-structured information in an acceptable amount of time.

4.1.6 What is missing?

The handling of complex noun phrases was a significant issue: On the one hand, the decisive context was lost if phrases were broken down into small entities. On the other hand, if phrases were retained in their original form, the context was kept, but the canonicalization remained unclear. To the best of our knowledge, there is no out-of-the-box solution that will solve these issues.

4.2 Pharmaceutical case study

We applied the toolbox to a subset of the biomedical Medline collection for our second case study. The PubMed Medline is available in different formats, among other things, in the PubTator format, which is supported by the toolbox. We downloaded the document abstracts from the PubTator Service [33].

4.2.1 Entity linking

We utilized existing entity annotations (diseases, genes, and species) from the PubTator Central service [33, 34]. In addition, we selected subsets of MeSH (diseases, methods, dosage forms), ChEMBL [24] (drugs and chemicals), and Wikidata [32] (plant families) to derive suitable entity vocabularies. We developed scripts that retrieved relevant entries from these vocabularies. This step required us to export relevant entries from XML and CSV files into TSV files.

We then applied the entity linker and analyzed the results by going through the most frequent annotations. Our first attempt yielded frequently, but obviously wrongly linked words such as *horse*, *target*, *compound*, *monitor*, and *iris*. These words were derived from ChEMBL because they were trade names for drugs. We found such trade names to be very ambiguous and removed them. Our final vocabulary included 69,502 distinct entities and 300,133 distinct terms.

But we also found annotations such as *major*, *solution*, *relief*, *cares*, *aim*, and *advances*. We went through the 500 most tagged entity annotations to remove such words by building a list of ignored words (188 in sum). We repeated the entity linking by ignoring these words and computed 232.5k entity mentions. We did not apply Stanford Stanza NER (persons, organizations, and more) here because we were interested in biomedical entities. The number of detected entities already seemed to be sufficient, so we continued with the extraction.

4.2.2 Information extraction

OpenIE6. The domain experts were interested in statements between entities. That is why we applied OpenIE6 and analyzed the partial and exact entity filter, i.e., we wanted to obtain entities as subjects and objects. We skipped no filter and subject filter here because they would have produced not-canonicalized noun phrases. OpenIE6 extracted 207.6k extractions and filtering them yielded 88k (partial) and 291 (exact) extractions. Our heuristic estimated 37.8% of the extracted subjects, and 72.1% of the objects as complex.

Table 5 PubMed PathIE example extractions. On the left, the canonicalized relation is annotated

Pharmacy	Treats	P1.1	We tested whether short-term, low-dose *treatment*[P] with the fluvastatin and **valsartan**[S] **(drug)** combination could improve impaired arterial wall characteristics in type 1 **diabetes mellitus**[O] **(disease)** patients
		P1.2	We encountered two cases of cerebellar **hemorrhage**[O] **(Disease)** in patients *treated*[P] with **edoxaban**[S] **(Drug)** for PVT after hepatobiliary surgery during the past 2 years
	Inhibits	P2.1	**Anthraquinone**[S] **(Drug)** derivative emodin inhibits tumor-associated angiogenesis through *inhibition*[P] of **extracellular signal-regulated kinase 1**[O] **(Gene)**/2 phosphorylation
		P2.2	Impact of **aspirin**[S] **(Drug)** on the gastrointestinal-sparing effects of **cyclooxygenase-2**[O] **(Gene)** *inhibitors*[P]
	Induces	P3.1	**Hyperglycemia**[O] **(Disease)**-*induced*[P] mitochondrial dysfunction plays a key role in the pathogenesis of diabetic **cardiomyopathy**[S] **(Disease)**
		P3.2	Conclusions H. pylori **Infection**[S] **(Disease)** appears to *cause*[P] decreases in **Vitamin B12**[O] **(Excipient)**[...]

Exact Entity Filter. The exact entity filter produced only 291 extractions out of 87.1k sentences (47.4k sentences with at least two entities). This method was hence too restrictive and not helpful because the remaining extractions were too few for a practical application.

Partial Entity Filter. A closer look at 100 randomly sampled extractions indicated that many noun phrases were complex again. The partial entity filter mixed up the original sentence information by filtering out the important information. For example, consider the following sentence: *Inhibition of P53-MDM2 interaction stabilizes P53 protein and activates P53 pathway*. Here the partial entity filter extracts the statement: (*MDM2*, *stabilizes*, *protein*). This statement mixed up the original information. Our analysis showed that the vast majority of filtered extractions were incorrect. In addition, OpenIE6 is focused on verb phrases to extract statements (here *stabilizes*).

However, many relevant statements are expressed by using special keywords, e.g., *treatment*, *inhibition*, *side effect*, and *metabolism*. That means that these OpenIE methods will usually not extract a statement from clauses like *metformin therapy in diabetic patients* by design. A similar observation was already made in the original toolbox paper, where OpenIE methods' recall was clearly behind supervised methods (5.8% vs. 86.2% and 6.2% vs. 75.9% on biomedical benchmarks) [15]. Supervised extraction methods would address this problem by learning typical patterns of how a treatment can be expressed within a sentence.

PathIE. To integrate such specialized keywords in the extraction process, we applied the recall-oriented PathIE method. In the previous example, the entities *metformin* and *diabetic patients* are connected via the keyword *therapy*. In this way, PathIE extracted a helpful statement. However, we had to build a relation vocabulary to define these specialized keywords. In cooperation with domain experts, we built such a vocabulary by incrementally extracting statements with PathIE, looking at extractions and example sentences to find out what we were missing. In sum, we had three two-hour sessions to build the final relation (eight relations plus 60 terms) vocabulary. The final PathIE step yielded 430.8k extractions and took two minutes to complete. Some interesting results are listed in Table 5. We then iterated over a sample of 100 of these extractions.

PathIE was capable of extracting statements from long and nested sentences, e.g., a treatment statement in P1.1 in Table 5. However, we also encountered several issues with PathIE. If a sentence contains information about treatments' side effects (also linked to diseases), PathIE extracted them wrongly as the treated condition (See P1.2). A similar problem occurred when a drug therapy was used to treat two diseases simultaneously. Here, PathIE yielded six statements (three mirrored): two therapy statements about the drug and each disease, and one therapy statement between both diseases, which is wrong. In example P2.2, PathIE failed to recognize that aspirin *effects* the inhibitors and is not an inhibitor itself.

A second problem was the direction of extracted relations: A *treats* relation could be defined as a relation between *drugs* and *diseases*. If a relation has precise and unique entity types, then an entity type filter can be used to remove all other, and possibly wrong, extractions. Suppose a disease causes another one (think about a disease that causes severe effects). In that case, PathIE would extract both directions: (a causes b) and (b causes a). For example, PathIE would extract two statements from *myocardial damage caused by ischemia-reperfusion*. Here an entity type filter did not solve the problem because both entities have the type *disease*.

Third, in situations with several entities and clauses within one sentence, PathIE seemed to mess up the original information and extracted wrong statements, e.g., see P3.1, where hyperglycemia did not induce cardiomyopathy. In summary, PathIE could extract statements from complex sentences, but a cleaning step had to be applied afterward to achieve acceptable quality.

4.2.3 Canonicalization

We exported the database statistics for PathIE. We carefully read the extracted verb phrases in cooperation with two domain experts. Verb phrases such as *treats*, *prevents*, and *cares* point toward a *treats* relation, which we included in our relation vocabulary. Phrases such as *inhibits* and *down regulates* may stand for a *inhibits* relation. To find more synonyms automatically, we used a Biomedical Word Embedding [38] that we used in our toolbox paper before. Following this procedure, we defined eight relations with 30 synonyms. We repeated the procedure five times and derived a relation vocabulary of 60 entries. The relation vocabulary was a mixture of verb phrases and keywords that indicated a relation in the text. In sum, we had six sessions of two hours each to build the final relation vocabulary.

However, we noticed that PathIE extractions were problematic when not filtered. Relations like *treats* and *inhibits* also include entity types that we had not expected, e.g., two diseases in treats. We formulated entity type constraints for eight relations to remove such problematic statements. The relations *treats* and *inhibits* looked more helpful because they only contained relevant entity types. We tried to filter relations like *induces* between diseases. Some extractions were correct, but many mixed up the relation's direction (a causes b instead of b causes a). In the end, PathIE was not very helpful for extracting such directed relations because of its poor quality. We stopped the cleaning here, but a more advanced cleaning would be helpful to handle such situations.

4.2.4 Application costs

We spent most of our time designing entity and relation vocabularies and analyzing the retrieved results. The creation of suitable vocabularies took us around one week in sum. The execution of the toolbox scripts was quite simple; see our GitHub repository. To measure the runtime for PubPharm, we applied the PathIE-based pipeline on around 12 million PubMed abstracts (PubMed subset about drugs). The procedure could be completed within one week: Entity detection took two days for the complete PubMed collection (33 million abstracts). PathIE took five days, and cleaning took one day. Hence, such an extraction workflow is realizable for PubPharm with moderate costs.

4.2.5 Generalizability

We already know that OpenIE6 and PathIE have worse performance than supervised methods; see the benchmarks in the original toolbox paper. However, we could design a suitable extraction workflow with an acceptable amount of time (a few weeks of cooperation with nine sessions with experts). OpenIE6 had a very poor recall, and filtering remained unclear. Thus, they were not of interest for PubPharm's purposes.

PubPharm is currently using the PathIE extractions in their narrative retrieval service[8] [16, 19]. Here recall is essential to find a suitable number of results to answer queries. Although the quality of PathIE is only moderate, the quality seems to be sufficient for such a retrieval service. Here, the statement should hint that the searched information is expressed within the document, e.g., that a *metformin treatment* is contained. The main advantage of a retrieval service is that the original sentences can be shown to users to explain where the statements were extracted. In summary, if users are integrated into the process, and the statements' origin is shown, PathIE allows novel applications like PubPharm's narrative retrieval service.

Nevertheless, we encountered several issues: First, PathIE extracted wrong statements if several entities were contained in a sentence. Next, the undirected extractions of PathIE were often problematic if no additional cleaning could be performed (e.g., relations between diseases). Although these issues must be faced somehow, PathIE allowed us an extraction workflow that we could not have realized using supervised methods due to the lack of training data. We would not recommend PathIE for building a knowledge graph because of many wrong extractions that would lead to transitive errors when performing reasoning on the resulting graph.

4.2.6 What is missing?

In this pharmaceutical case study, we focused on relations between pharmaceutical entities. PathIE completely ignored the surrounding context of statements, e.g., dose and duration information of therapies. The coherence of statements was also broken down, e.g., drug, dosage form, disease, and target group of treatments were split into four separate statements. The desired goal would be to retain all relevant information within a single statement. However, PathIE is restricted to binary relations. A future enhancement of PathIE would be desirable to retain all connected entities in a sentence. PubPharm's narrative retrieval service bypassed the problem by using document contexts [18], i.e., statements from the same document belong together. The service used abstracts, and this approximation would not have been possible for full texts

[8] www.narrative.pubpharm.de.

Table 6 Pollux OpenIE6 example extractions. On the left, the corresponding entity filter is shown (partial and subject)

Political sciences	Partial	PS1.1	Stalin wanted all 16 **Soviet**[S] **(NORP)** Republics *to have*[P] separate seats in **UN General Assembly**[O] **(ORG)** but only 3 were given Russia Ukraine Belarus
		PS1.2	This paper seeks to understand why the **United States**[S] **(GPE)** *treated*[P] **Japan**[O] **(GPE)** and Korea *differently*[P] in the revisions of bilateral nuclear cooperation agreements
	Subject	PS2.1	Based on these features, the article suggests that **China**[S] **(GPE)** is poised *to become*[P] **a true global power**[O]
		PS2.2	Prior to the introduction of the Transparency Register the **European Parliament**[S] **(ORG)** *had maintained*[P] **a Register of Accredited Lobbyists since 1996**[O] while the European Commission [...]

because a full-text document might contain several different contexts.

4.3 Political sciences

We applied the toolbox to 10k abstracts from Political Sciences.

4.3.1 Entity linking

The field of Political Sciences displays some distinct differences compared to the biomedical field and encyclopedias like Wikipedia. A notable difficulty lies in the lack of well-curated vocabularies for the domain. This can be mitigated in two ways: by using NER as implemented by Stanza [27] or by constructing/deriving entity vocabularies from general-purpose knowledge bases like Wikidata. We investigated both approaches.

Stanza NER yielded ca. eight tags per document. The extracted mentions seemed sensible, e.g., entities like *USA*, *Bush*, or *the Cold War* were extracted. Problematic was that mentions like *Bush* were identified as a *person* and not linked to a specific identifier. However, Stanza NER also displayed some drawbacks, e.g., it was prone to missing uppercase letters for identifying names. Such restrictions can be problematic in practice because of bad metadata, e.g., abstracts in upper case.

For the second approach, we selected wars (Q198), coup d'états (Q45382), and elections (Q40231) as seed events, since those are likely to be the subject of debate in political science articles. Furthermore, we inductively utilized Wikidata's subclass property (P279) to receive all subclasses of all seed events. We used the SPARQL endpoint to export the corresponding vocabularies by asking for the English label and alias labels for the seed events, all instances of the seed events (P31– instance of), and their subclasses. In total, we collected 2.9k wars, 904 coups, and 79.7k election entries. An evaluation of the toolbox's entity linker showed good performance on wars, while coup d'états and elections were rarely linked sensibly. Our vocabulary included 52,454 distinct entities and 59,813 distinct terms.

However, we increased the linking quality by applying simple rules, e.g., the entity label must contain the term *election*. We derived 3.7k entity annotations linked to Wikidata in sum.

4.3.2 Information extraction

OpenIE6. Due to the lack of comprehensive entity vocabularies, we focused on OpenIE6 in this case study and omitted PathIE. OpenIE6 yielded 147.2k (no filter), 28.6k (partial), 128 (exact) and 7.3k (subject) extractions. Subject phrases tended to be short (only 32.0% were complex), and object phrases tended to be long (74.3% complex) again, like in the previous case studies. We randomly sampled 100 extractions of each filter for further analysis. Again, extractions from small sentences looked helpful, while long sentences led to long object phrases. We picked some interesting results and displayed them in Table 6.

Exact entity filter. Again, the exact entity filter decreased the number of extractions drastically (from 147.2k to 128). But extractions seemed plausible, e.g., *Alexander Lukashenko is president of Belarussian[SIC]* from *Focus on the career and policies of the first Belarussian president, Alexander Lukashenko, elected in 1994.* Another correct extraction was *United States prepares to exit* from *As the United States prepares to exit Afghanistan [...].*

Partial entity filter. In PS1.1, the extraction *Soviet to have UN General Assembly* was wrong because the context about Stalin and separate seats was missed. The extraction in PS1.2, *United States treated differently Japan*, was not helpful because *Korea* was missed. Again, the context that this statement was investigated in that article was lost. We found the extractions of the partial filter not helpful: Either they mixed up the original information, or decisive context was missed.

Subject entity filter. The extraction PS2.1 showed a correct extraction, but then the information that the statement was suggested by an article was missed. Although the sentence of PS2.2 was quite complex, OpenIE6 extracted useful information about the *European Parliament*: *European Parliament had maintained a Register of Accredited Lobbyists since 1996*.

4.3.3 Canonicalization

We exported the most extracted verb phrases and analyzed them. The ten most frequently extracted verb phrases (lemmatized) were: be, have, be in, provide, examine, present, offer, focus on, be with, and may. We skipped the canonicalization procedure here because we already knew that canonicalizing OpenIE6 verb phrases remains unclear (see Wikipedia case study). The more so, when words like *be, provide, offer* or *may* could refer to various relations—again depending on the context.

The exact filter yielded fewer extractions, partial filtering resulted in incorrect statements, and PathIE could not be applied due to the lack of vocabularies. And extractions from the subject filter could hardly be canonicalized to precise relations if the object phrase contained large sentence parts (complex object noun phrases).

4.3.4 Application costs

The application costs for the political domain seemed higher compared to the other two case studies. The lack of curated vocabularies necessitates the creation of such. As demonstrated, this can hardly be done automatically but requires domain knowledge. We exported some vocabularies from Wikidata but missed many entities in the end. In sum, we had four sessions, each 1.5 h, with a domain expert to analyze the results. The case study took us five person-days in sum.

4.3.5 Generalizability

Due to the lack of available benchmarks, we restricted our evaluation to a qualitative level. As another difficulty, simple fact statements, e.g., *Joe Biden is the president of the USA* hardly carried new or relevant information. Still disputed claims, viewpoints, or assessments like *the UK aims to position itself as an independent power after Brexit* might be the subject of study. This often resulted in long clauses for the subjects and objects that are hard to map to the already sparsely recognized named entities. But the subject entity filter allowed us to retain that *UK aims to position itself as an independent power after Brexit* as a suitable extraction. We plan to proceed from here by extracting semi-structured information via the subject filter.

4.3.6 What is missing?

Additionally, the context of a statement is often highly relevant. In the example, the statement loses its information if the context *after Brexit* is omitted. Observations were similar to the Wikipedia case studies: Either the object phrases retained the context but could hardly be handled by filtering methods. Or the object phrases were short and missed information.

4.4 On complex noun phrases

In the following, we use different methods to analyze the complexity of OpenIE noun phrases in more depth. We then continue by looking at the CoreNLP OpenIE extraction tool to generalize our previous findings better, especially, if they are just an artifact of OpenIE6. All implemented extensions, developed scripts, and produced and analyzed data can be found in our repository.[9]

In the previous case studies, we used a self-developed heuristic to estimate if an OpenIE noun phrase is complex. The heuristic was based on information about the length of the noun phrase, whether the sentence has multiple clauses and a few regular expressions. In the following, we applied a bunch of different methods to analyze the complexity of noun phrases in more detail.

Basically, our methods can be grouped into two categories: (1) Part-of-Speech (POS) tag-based and (2) character length-based methods. A POS tag aligns a word of a sentence to a certain part of speech, e.g., nouns, pronouns, adjectives, and more. The evaluation here was based on utilizing such POS tags. For example, we analyzed how many noun phrases contained verbs. Therefore, we used the Universal POS tags.[10] We classified whether an OpenIE noun phrase felt into one of the following categories:

1. Has an adposition (ADP),
2. Has a conjunction (CCONJ),
3. Has nouns only (NOUN, PROPN, PART, DET, NUM, PUNCT),
4. Has nouns and pronouns only (same as for nouns + PRON),
5. Has nouns, pronouns, and adjectives only (same as for nouns + PRON + ADJ), and
6. Has a verb (VERB).

Our motivation for complex noun phrases was that they should include more than a single concept, e.g., a whole sentence fragment or a composition of concepts. That is why we analyzed adpositions to count how many noun phrases

[9] https://github.com/HermannKroll/KGExtractionToolbox/blob/main/README_IJDL2023.md.

[10] https://universaldependencies.org/u/pos/.

contain words like *of*, *in*, *during*, etc., which may indicate a composition of concepts. We also counted conjunctions for the same reason. We also focused on nouns, i.e., we counted how many noun phrases only consisted of nouns. Note that we allowed the following tags for nouns: PROPN to also allow proper nouns, PART to allow fragments like ' in nouns, DET to allow words like *the, a, an, etc.*, NUM to allow numbers (e.g., 3 cats) and PUNCT to allow abbreviations (e.g., St. Paul). In two additional categories, we also allowed nouns and pronouns as well as nouns, pronouns, and adjectives. For comparison, we also counted verbs in noun phrases, which may indicate a relation between concepts. In brief, we understand a noun phrase consisting of nouns only as not being complex. Conjunctions, adpositions, or verbs in noun phrases may likely hint toward a complex concept. To derive POS annotation, we applied the NLP Spacy[11] tool in version 3.1.4. We downloaded the English model (*en_core_web_sm*) for our subsequent analysis.

The second evaluation category was based on character length. The motivation was to understand better the ratio between the length of a noun phrase and the overall sentence length. We assumed long noun phrases to be complex, especially if they were longer than half of the sentence's length, for example. Therefore, we computed the length for each noun phrase and each sentence by counting the corresponding characters. Hence, we counted how many noun phrases were longer than 30%, 40%, 50%, 60%, and 70% of the sentence.

4.4.1 Results

The evaluation results of our noun phrases extracted by OpenIE6 are reported in Table 7. First, extracted subjects were less complex for all methods and all domains. This reflected our previous findings that OpenIE6 subjects seemed less complex. And that objects were rather often complex. For example, 84.2% of all OpenIE6 subjects extracted from Wikipedia consisted of nouns, pronouns, and adjectives only. In other words, 15.8% subjects were thus more *complex* than a single noun. Our initial heuristic estimated 16.2% of the Wikipedia subjects to be complex. This argument also applies to Pharmacy and Political Sciences. Our heuristic estimated around 37.8% (Pharm.) and 32.1% (Pol.) to be complex. The noun+pronoun+adjective estimation revealed that around 42.1% (Pharm.) and 34.4% (Pol.) contained more information than a single noun. Broadly a third of all OpenIE6 objects in all three domains contained a verb. Concerning the noun phrase length, between 25.5 and 28.6% of the objects were longer than 40% of the sentence. Indeed, between 15.1 and 18.2% were longer than 50% of the sentence. This quantified our qualitative impression that many extracted noun phrases consisted of whole sentence fragments.

To better generalize our findings here, we applied another OpenIE tool, namely CoreNLP OpenIE on our data. This method is older (2014) than OpenIE6 (2020) and may likely have different properties. After execution, we obtained 545k extractions for Wikipedia, 930k for PubMed, and 569k for Political Sciences. The first observation was that CoreNLP OpenIE extracted way more statements than OpenIE6 (545k vs. 179k, 930k vs. 210k, and 569k vs. 150.7k). A quick investigation revealed that CoreNLP OpenIE extracted several *similar* statements from sentences, e.g., five extractions from *The quick brown fox jumped over the lazy dog*. Here, the tool extracted three different versions of the subject (quick brown fox, brown fox, fox), the verb phrase *jumped over*, and the two objects (lazy dog and dog) – yielding six extractions in sum. For this example, OpenIE6 extracted a single extraction: (The quick brown fox; jumped; over the lazy dog). Additional filtering might be beneficial here. However, how to do so is challenging, e.g., keeping just the longest extraction in terms of noun phrase length may conflict with the exact or subject entity filtering later on. If the fox was an entity and we just kept the quick brown fox as the only subject, our filtering methods would not produce a result here. But keep this property in mind for the following investigations.

The noun phrase complexity of CoreNLP OpenIE is reported in Table 7. In brief, this method extracted less complex noun phrases for subjects and objects for all three domains measured by our heuristic. A closer look at the other estimation methods revealed that those supported our findings. The ratios of pure noun phrases consisting of nouns or nouns+adjectives+pronouns were clearly above the ratios of OpenIE6.

In our previous manual evaluation [14], we manually counted the complexity of noun phrases for biomedical and new articles. The findings back then revealed that between 53 (biomedicine) and 68% (news) of OpenIE6 extracted objects were classified as complex by raters. For CoreNLP OpenIE, in contrast, we estimated 25% (biomedicine) and 20% (news) as complex objects. Concluding from both findings (this paper) and our previous study [14], the main takeaway here is that complex noun phrases are a frequent issue that must be faced in practice. Although less frequent for CoreNLP OpenIE than for OpenIE6, they are still there. Handling such complex noun phrases by canonicalizing methods like entity filters still remains open.

4.5 CoreNLP OpenIE

In the first case study, we investigated the noun phrase complexity of CoreNLP OpenIE in comparison to OpenIE6. Although the tool seemed to have less noun phrase complexity, how useful are its extractions in practice? First, we

[11] https://spacy.io/.

Table 7 Evaluation of the OpenIE noun phrase complexity: Different methods and features are used to estimate how *complex* an OpenIE noun phrase is. Therefore, the method, its features, and the results for subjects and objects, as well as for all three domains, are reported. Note that the OpenIE6 complexity is based on 179k tuples for Wikipedia, 210k for PubMed, and 150.7k for Political Sciences. For CoreNLP OpenIE, the results are based on 545k tuples for Wikipedia, 930k for PubMed, and 569k for Political Sciences

Method	Features	Wikipedia		Pharmacy		Pol. sciences	
		Subj. (%)	Obj. (%)	Subj. (%)	Obj. (%)	Subj. (%)	Obj. (%)
OpenIE6							
Our heuristic	Mixed	16.2	74.5	37.8	72.1	32.1	74.4
Has adposition	POS Tags	10.5	77.3	30.4	79.6	24.8	76.3
Has conjunction	POS Tags	0.3	2.3	1.7	4.5	1.9	6.0
Has nouns only	POS Tags	43.0	9.2	32.5	6.3	37.9	7.2
Has nouns+pronouns only	POS Tags	76.0	10.6	40.9	6.5	50.3	7.8
Has n.+pron.+adj. only	POS Tags	84.2	15.1	57.9	12.0	65.6	13.3
Has verb	POS Tags	5.7	29.3	16.7	33.4	13.1	36.9
> 30%-of-Sentence	Char. Length	4.3	43.1	14.1	40.4	10.8	41.9
> 40%-of-Sentence	Char. Length	1.7	28.6	6.7	25.5	5.0	27.3
> 50%-of-Sentence	Char. Length	0.7	18.2	2.9	15.1	2.1	17.0
> 60%-of-Sentence	Char. Length	0.2	11.4	1.2	8.5	0.8	10.1
> 70%-of-Sentence	Char. Length	< 0.1	6.5	0.4	4.0	0.3	5.3
CoreNLP OpenIE							
Our heuristic	Mixed	3.1	46.8	4.0	53.0	2.8	53.0
Has adposition	POS Tags	0.3	45.4	0.9	52.1	0.3	52.6
Has conjunction	POS Tags	0.1	0.1	< 0.1	0.1	< 0.1	< 0.1
Has nouns only	POS Tags	45.2	28.0	50.8	19.3	53.6	20.3
Has nouns+pronouns only	POS Tags	80.1	31.1	59.5	19.6	66.6	21.4
Has n.+pron.+adj. only	POS Tags	91.2	41.3	82.3	31.5	88.2	33.1
Has verb	POS Tags	7.2	32.4	15.3	40.3	10.8	37.5
> 30%-of-Sentence	Char. Length	0.5	19.5	1.1	22.8	0.8	20.3
> 40%-of-Sentence	Char. Length	0.1	10.6	0.2	11.6	0.2	10.6
> 50%-of-Sentence	Char. Length	< 0.1	5.1	< 0.1	4.9	< 0.1	4.8
> 60%-of-Sentence	Char. Length	< 0.1	2.1	< 0.1	1.6	< 0.1	1.7
> 70%-of-Sentence	Char. Length	< 0.1	0.6	< 0.1	0.3	< 0.1	0.4

had a close look at existing NLP benchmarks [3, 6, 11]. In brief, OpenIE6 outperformed CoreNLP OpenIE. We made a similar observation when quantifying how much information these tools keep in practice; see [14]. These findings were expected because the CoreNLP OpenIE is way older and less advanced than OpenIE6.

However, our entity filtering approaches have revealed that handling complex noun phrases remained unclear because either the exact filter yielded too less extractions in practice, or the partial filter mixed up the original sentence's information. Due to a less noun phrase complexity when using CoreNLP OpenIE, we formulated the questions: *1. Does the partial entity filter obtain a better overall quality? 2. Does the exact entity filter yield a sufficient number of extractions in practice? 3. Should we switch back to CoreNLP in combination with entity filtering?*

Table 8 We report the number of CoreNLP OpenIE extractions computed by the different entity filters (no, partial, exact, subject) for our three domains

CoreNLP OpenIE				
Ent. Filter	*#No*	*#Part.*	*#Exact*	*#Subj.*
Wikipedia	544k	171k	36k	272k
Pharmacy	929k	466k	7.7k	112k
Pol. Sci.	568k	11.2k	1.2k	30k

4.5.1 Extraction and filtering

We applied the CoreNLP OpenIE method to our previous case study data by using the same entity annotations for filtering as we used for OpenIE6. The resulting numbers of extractions for each entity filter (no, partial, exact, and subject) are reported in Table 8.

First, the overall number of extractions without filtering was higher than using OpenIE6. We commented on this finding in the previous subsection. For the exact filter, the number of remaining extractions was higher than in the OpenIE6 setting: This time we obtained 36k, 7.7k, and 1.2k extractions instead of 2.9k, 291, and 128 extractions. However, how useful were these extractions? So, we (two authors) performed a qualitative evaluation of the filtered results. We randomly sample 50 extractions for each filter (partial, exact, and subject) and each domain, i.e., 450 in total.

4.5.2 Wikipedia

Partial Filter. The results of this filter were similar to our findings for OpenIE6. We saw some good extractions like (Dahleh, is, professor) from *Munther A. Dahleh [...] is the William Coolidge Professor [...].* However, we also saw many situations in which the partial filter mixed up the original information, e.g., (Birkeland, was born to, Birkeland) from *Birkeland was born in Christiania (Oslo today) to Reinart Birkeland and Ingeborg [...]*, or (Alexander von Humboldt, is, German) from *Alexander von Humboldt is also a German ship named after the scientist [...].*

Exact Filter. Although the exact filter yielded better extractions, the question was how useful were the extractions in the end. Suppose the following three examples: 1. (Schuenemeyer, is president of, Colorado) from *Schuenemeyer is President of Southwest Statistical Consulting, Cortez, Colorado*. 2. (Niebur, was, president) from *Niebur [...] was president of the National Association of Graduate*. 3. (Wegelin, succeeded langhans as, director) from *[...] Wegelin succeeded Langhans as director of the Anatomical institute*. In all cases, the extraction was syntactically correct. However, the extractions were not useful. Schuenemeyer is not the president of the state of Colorado. He is the president of an organization in Colorado. The organization/affiliation of Niebur's presidency was missed, too. Wegelin indeed succeeded Langhans as a director, but in which position?

Subject Filter. This filter yielded the original object phrases that were extracted by CoreNLP OpenIE, e.g., (Faruque, maintained, active research team) from *Faruque maintained an active research team in icddr [...]*, or (Thoguluva Shesadri Chandrasekar, is, Indian gastroenterologist) from *Gastroenterologist Thoguluva Shesadri Chandrasekar (born 1956) is an Indian gastroenterologist [...].* However, we found that these object phrases were shorter than for OpenIE6, and hence, did contain less information.

4.5.3 Pubmed

Partial Filter. Similarly to the Wikipedia findings, we found it hard to evaluate extractions like (Patients, is with, Disease) from *We identified 8 patients (7 with ALS and 1 with SMA) with motor neuron disease [...].* Although the extraction might be rated as correct, it was not very helpful. The information about the number of patients and which concrete disease was missed. Another extraction was (Injection site reactions, were considered by, Patients) from *Local injection site reactions, including swelling [...], were considered mild or moderate by the patients [...].* The extraction missed how the reactions were considered. So is it correct? Likely yes, but useless.

Exact Filter. (Granulomas, presence of, lymphadenopathy) from *Years later, the presence of pathologic submandibular lymphadenopathy was identified and biopsied, revealing non-caseating granulomas* was a wrong extraction. In contrast, the following three extractions looked correct: 1. (Preterm birth, is contributor to, infant death) from *Preterm birth (PTB) is the largest contributor to infant death in sub-Saharan Africa [...].* 2. (abpa, is usually associated with, respiratory diseases) from *Allergic bronchopulmonary aspergillosis (ABPA) [...] is usually associated with underlying respiratory diseases such as asthma or cystic fibrosis.* 3. (Testicular cancer, affect, men) from *Testicular cancer and Hodgkin's disease are among the most common malignancies to affect young men of reproductive age.* However, much information was still lost: Which men are affected?

4.5.4 Pollux

Partial Filter. Again, the partial filter was problematic, e.g., consider the extraction (Woodward, once again pulls back, Washington) from *Woodward once again pulls back the curtain on Washington [...].* Alternatively, consider: (Switzerland, member of, UN) from *Prior to its full membership in the United Nations, Switzerland was an active observer and even an active member of many specialized UN agencies.* The first extraction missed what was pulled back, and the second one was problematic, too: Here, Switzerland was a member of specialized UN agencies. So UN was detected as an entity, but the rest was missed.

Exact Filter. Extractions like (Putin, is more isolated after, nearly a decade) from *After nearly a decade in power, Putin is more isolated than ever* looked syntactically correct. Another one was (Chaldeans, is in, Iraq) from *[...] experienced by the Chaldeans in Iraq in the last two decades.* We observed many (s, *is in*, o) extractions based on the word *in*. In addition, we also observed problems with '-based extractions like (Nkrumah, of, Ghana) from *The Case of Nkrumah's Ghana.*

Subject Filter. Analogous to our previous observations, the subject filter yielded results of mixed quality. We observed extractions like (South African Defence force, facilitated, relocation of about 4000 bushmen from military bases) from *In March 1990 the now defunct South African Defence Force facilitated the relocation of about 4000 bushmen from military bases [...]* which correctly repeated the gist of the

original sentence but omitted context information (e.g., when the relocation happened and that the defence force was defunct). Yet again, short object phrases can lead to rather useless extractions, e.g., (Spain, is second most important country in, terms) from *In the case of Wind Energy, and in terms of production, Spain is the second most important country […]*.

4.5.5 Results

We observed that CoreNLP OpenIE indeed extracted less complex noun phrases than OpenIE6. However, these less complex noun phrases also mean that less context and coherence of the sentence was kept. The partial filter still mixed up with the original information or broke down information into pieces. The exact filter retrieved a higher number of extraction, but the overall quality seemed to be lower than in the OpenIE6 setting, likely because the CoreNLP tool itself had a lower extraction quality. The subject filter still seemed to work out: Subjects were linked to entities, and objects remained not filtered. However, we would still recommend using OpenIE6 for subject filtering. On the one hand, OpenIE6 had a better overall extraction quality (see NLP benchmarks). On the other hand, OpenIE6 extracted longer noun phrases as objects, i.e., more information is kept in that objects. The key takeaway here was that our previous findings for OpenIE6 also applied for OpenIE, allowing a better generalization of our overall findings.

4.6 A remark on quantification

A good question is why our evaluation was mainly qualitative instead of quantitative in nature. On the one hand, existing NLP benchmarks already report on the pure extraction quality and, likely, have a better quality than we would achieve. On the other hand, our goal was to discuss the challenges of information extraction workflows in digital libraries. For example, although the extraction (Patients, is with, Disease) might be seen as syntactically correct, it still does not seem useful in practice. And even worse, our workflow relied on the quality of entity detection, information extraction, filtering and canonicalization, so that each step might lead to subsequent errors. As an example, we quantified the CoreNLP OpenIE extractions of Wikipedia for the partial filter. We would rate 17 of 50 as correct. However, twelve of them were about persons, and six of them had wrongly identified entities. And even worse, some of the correct ones had only partial person names tagged, so just Einstein or Turing, instead of their full names. For the exact filter on Wikidata, we would rate 43 of 50 as correct— but 21 of them had wrongly linked entity types (Washington as a location instead of a person). In the end, we found the quantification too challenging, and the resulting numbers could still be wrong and hence, misleading in the end. That is why we focused on a qualitative study to show the opportunities and drawbacks of such inf. extraction workflows.

5 Advanced canonicalization

Our initial verb phrase canonicalization approach was based on designing a relation vocabulary, i.e., define relations plus a set of synonyms. Such a design can be challenging, as our case studies showed. Canonicalizing verb phrases without considering their sentence contexts remained unclear. Subsequently, we discuss another verb phrase canonicalization based on clustering.

Vashishth proposed CESI to canonicalize OpenIE extractions by clustering noun and verb phases with the help of side information [31]. We wanted to investigate how useful this idea is in practice, i.e., clustering verb phrases that would not require the design of a relation vocabulary. Therefore, we implemented an additional canonicalization method into our toolbox that works as follows: 1. All verb phrases of the extractions are retrieved. 2. These verb phrases are embedded by word embedding that must be given as input. 3. Clustering is performed, and the results are shown to the user.

However, by implementing the last step, we followed the procedure of CESI.[12] They used agglomerative clustering to bypass the need for a pre-given number of clusters. However, a threshold must be provided for splitting the actual clusters. And especially this threshold caused issues for us: How to select a *suitable* threshold?

Here, we used the same Wikipedia Word Embedding as in our case studies before. And, we used the OpenIE6 extractions again. Using the default threshold of 0.429 (see CESI implementation) yielded 351 clusters for 1062 distinct verb phrases from Wikipedia. One cluster, for example, contained the verbs *stand* and *sit*. Another cluster contained the verb phrases *be take, take over, to take, take up, take on, have take, have take over, to take up*. One cluster even contained 629 different verb phrases. We obtained 380 clusters for distinct 1145 different verb phrases from our Political Sciences sample. Alternatively, a threshold of 0.5 yielded 150 and 165 clusters. A threshold of 0.6 yielded 20 and 33.

First, verb phrases need eventually be better cleaned (removing words like *be, to, up, on, by, etc.*) for a practical application. Second, selecting a suitable threshold is challenging. In the end, such a clustering approach did not solve the overall problem that we faced in our case studies. Verb phrases like *use* require the sentence's context information to be reliably canonicalized because they could refer to many different relations. However, such a clustering might give first ideas of which relations could be hidden in the text. So

[12] https://github.com/malllabiisc/cesi/blob/master/src/cluster.py.

it could be used to create a relation vocabulary. Nevertheless, we already developed a script in the original toolbox to export which verb phrases appear most frequently across the collection.

6 Non-English texts

Digital libraries cover a large quantity of texts in different languages. This is especially true for national libraries, e.g., the German National Library or the Royal Library of the Netherlands. In such cases, there is a need for information extraction tools supporting those languages. However, besides some notable exceptions (CoreNLP), most tools are not capable of dealing with non-English texts. They are thus limited in usage for such cases. This is because, besides huge advances in natural language processing in the last decade, there is a clear lack of research in this area regarding texts in languages other than English; see [4] for a good discussion. Thus, other solutions are needed to adapt to non-English texts.

One solution for a couple of languages might be to utilize machine translation for the documents. There is work in the direction of translating training data to train OpenIE systems for other languages [12]. Our idea here was to translate the non-English text into English and apply the toolbox on top of the translation. This approach did not require to adjust the actual methods or retrain NLP models. And, if possible, it would allow utilizing the toolbox's methods on a larger variety of languages since modern machine translation systems support a myriad of languages. That is why we investigated if we can handle Non-English texts (here: German texts) by using automated machine translation. According to this idea, we formulated our research question:

Could machine translation be a solution to handle non-native English texts? And if, how well does the workflow apply here?

6.1 Content

For this small case study, we again focused on the previous three domains: Wikipedia, Pharmacy, and Political Sciences. We manually selected the Wikipedia articles of five famous scientists (Albert Einstein, Alan Turing, Max Weber, Sir. Roger Penrose, and Fritz Jakob Haber). We downloaded the English and German abstracts of these articles. We used the English abstracts for comparison, i.e., the basic idea was to compare sentences from the original English article and from the German-to-English translated one that contain a *similar* information. We were aware that Wikipedia articles might have different levels of detail in different languages. For Pharmacy, we asked a domain expert to provide us with ten pharmaceutical articles that contain an English and a German abstract. We downloaded four articles from *Krankenhauspharmazie*, three from *Phytotherapie*, and three from *Die Pharmazie*. For Political Sciences, we randomly sampled ten articles from the Pollux dump that contained an English and a German abstract. We used the English abstract for Pharmacy and Political Sciences to compare the extractions. The articles should—at best—contain the same information in both languages, i.e., the German-to-English translated version should be similar to the actual English hand-written version.

6.2 Translation service

For the translation, we used the known online service DeepL.[13] DeepL is free-to-use for documents up to 5,000 characters. Additionally, it offers a simple online API and can be adapted for practical scenarios. Note that the English, German, and German-To-English translated abstracts are available in our toolbox repository.

6.3 Statistics

We applied the same extraction workflow as we did for our main case studies, i.e., we used the same entity vocabularies as we used for the corresponding domain in our OpenIE6 case study. We did not adjust any vocabulary for this investigation.

Statistics about this case study's data are listed in Table 9. The Wikipedia articles contained 82 sentences, whereas the German-to-English translated version only contained 55 articles. For Pharmacy, the original English articles contained 14 sentences more, and for Political Sciences, the difference was three. For Wikipedia, 58 of 82 (70%) English sentences contained two entities comparable to the translated version, whereas 37 of 55 (67%) sentences contained at least two entities. The reason might be the different levels of detail in the English and German articles.

For Pharmacy, the number of sentences was decreased by 19%, the number of sentences with two entities by 16%, and the number of detected entities by 21%. For Political Sciences, the numbers of sentences with two entities, NER tags and EL tags were equal except for an entity linking problem: DeepL translated a German fragment to 1980s and 1990s, which were wrongly linked to a plethora of different Wikidata entities: 421 wrong links in total. For the subsequent analysis, we applied the same workflow as in our previous case studies, i.e., applied OpenIE6 with the no filter option; see Table 10 for statistics.

For the subsequent qualitative analysis, we (two authors) evaluated the pure OpenIE6 extractions (i.e., no filtering) to analyze how much information is kept from the original German sentences and how these extractions compare to the original English version. Table 11 shows a comparison of

[13] http://deepl.com.

Table 9 Statistics of our Non-English Case-Study. The numbers of sentences (#Sent.), sentences with at least two detected entities (#with2E), and the number of NER and EL tags are shown. T. denotes the German-to-English translations. *Note that 421 are wrongly linked entities

	Sentences		Entity Det.	
	#Sent.	*#with2E*	*#NER*	*#EL*
Wiki.	82	58	157	143
Wiki. T.	55	37	86	78
Pharm.	89	44	–	147
Pharm. T.	75	38	–	121
Pol. S.	70	11	27	3
Pol. S. T.	67	11	27	424*

Table 10 Translation Case Study: We report the number of extractions obtained from applying OpenIE6 with different entity filters (no, partial, exact, subject)

OpenIE6				
Ent. Filter	*#No*	*#Part.*	*#Exact*	*#Subj.*
Wikipedia	229	200	4	71
Wikipedia T.	119	78	–	17
Pharmacy	201	66	–	10
Pharmacy T.	180	68	–	8
Pol. Sci.	161	6	–	11
Pol. Sci. T.	167	4	–	8

OpenIE6 extractions from English and German-to-English translated texts. In addition, we show the original German texts.

6.4 Wikipedia

Again, remember that Wikipedia abstracts may differ in the levels of detail between the English and German versions. First, extracting information from the first sentence of Wikipedia, the description of who the scientist was, usually worked very well. For example, the extracted statements for Albert Einstein only differed by the word *theoretical* in the object because it was not mentioned in the German text. We made a similar observation for the other four scientists: If their descriptions were the same in English and German, then the translated version resulted in the same statements. Small derivations like that Alan Turing was described as a mathematician, philosopher, computer scientist, logician, and theoretical biologist in the English Wikipedia. In contrast, the translated text yielded the extraction that Turing was a logician, mathematician, cryptanalyst, and computer scientist.

Another sentence about the famous work of Einstein, see Table 11, yielded that Einstein is known for developing the theory of relativity from the English Wikipedia. In the German version, however, the information was stated in a nested version, i.e., the translated version was: *Einstein's main work, the theory of relativity, [...]* which did not yield a statement that he is known for his theory. So small changes in the formulation were decisive in whether a statement was extracted.

Another interesting finding was about Max Weber's occidental rationalism and the disenchantment of the world. The translation for this sentence worked very well, but the final extraction then yielded different statements than the English version, mainly because the formulation was quite different. However, that his work was developed by the unity of a leitmotif was still correctly extracted.

The German statement about Albert Einstein: *Für seine Verdienste um die Theoretische Physik, [...], erhielt er den Nobelpreis des Jahres 1921, der ihm 1922 überreicht wurde* was well translated into English: *For his services to theoretical physics, [...], he was awarded the Nobel Prize of 1921, which was presented to him in 1922.* OpenIE6 yielded the correct extraction that *he received the 1921 Nobel Prize*. The English article stated that *He received the 1921 Nobel Prize in Physics [...].* For this sentence, the extraction also contained the information that *he received the 1921 Nobel Prize* **in Physics**.

In brief, we observed many well-translated sentences and hence, many extractions that were comparable to the original English version, except for minor changes between the different articles.

6.5 Pharmacy

The first statement (see Table 11) about the head and neck region tumors yielded similar extractions except for some slight formulation derivations. In particular, the domain-specific terms were well translated here. This was also reflected by the number of detected entities which was quite close between the English and the German-to-English translated versions.

An interesting finding was the second statement. Although the German sentence was well translated and close to the original English version, OpenIE6 extracted two statements for the English version and only one for the translated version. The difference was based on a missing comma in front of the last *and* in the translated sentence. We manually added the comma and OpenIE6 yielded two statements again. Another finding was about formulations in the articles. The English abstracts tended to use the active formulation *we show*, whereas the German abstracts, and hence, the translated version tended to use the passive style like *it has been shown*. OpenIE6 extracted statements from the active version, but not from the passive version.

Overall, we observed many useful extractions from the translated version, and these extractions were close—except for some formulations—to the original English extractions.

Table 11 Comparison of OpenIE6 extractions from English and German-to-English translated texts

English	German-To-English Trans	German
Wikipedia		
Albert Einstein[S] was a **German-born theoretical physicist**[O]	**Albert Einstein**[S] was a **German-born physicist**[O]	Albert Einstein war ein gebürtiger deutscher Physiker
Einstein[S] is *best known*[P] **for developing the theory of relativity**[O]	**Einstein's main work**[S], the theory of relativity, *made*[P] **him world famous**[O]	Einsteins Hauptwerk, die Relativitätstheorie, machte ihn weltberühmt
Weber's main intellectual concern[S] *was*[P] in **understanding the processes of rationalisation**[O], secularisation, and the ensuing sense of "disenchantment	Even though his work is fragmentary in character, **it**[S] *was nevertheless developed*[P] **from the unity of a leitmotif**[O]: occidental rationalism and the disenchantment of the world it brought about	Auch wenn sein Werk fragmentarischen Charakter hat, wurde es dennoch aus der Einheit eines Leitmotivs entwickelt: des okzidentalen Rationalismus und der damit bewirkten Entzauberung der Welt
Pharmacy		
Tumors in the head[S] and neck region *include*[P] a **heterogeneous group of carcinomas whose treatment has advanced in recent years**[O]	**Tumors in the head**[S] and neck region *comprise*[P] a **heterogeneous group of carcinomas for whose therapy progress has been observed in recent years**[O]	Tumoren im Kopf-Hals-Bereich umfassen eine heterogene Gruppe von Karzinomen, für deren Therapie in den letzten Jahren Fortschritte beobachtet werden konnten
This work[S] *focuses*[P] **on radiation therapy**[O], a treatment option with possible short- and long-term complications, and **the resulting consequences for the patients' quality of life**[O]	**The focus here**[S] *will be*[P] **on radiation treatment**[O], a treatment option with potential short- and long-term complications and the resulting consequences for patients' quality of life	Im Vordergrund soll hier die Strahlenbehandlung stehen, eine Behandlungsoption mit möglichen kurz- und langfristig auftretenden Komplikationen sowie den daraus folgenden Konsequenzen für die Lebensqualität der Patienten
Political sciences		
Beginning with the **Mont Pelerin Society**[S], *founded*[P] by **the Austrian economist and philosopher Friedrich v. Hayek in 1947**[O], [...]	Starting with the **Mont Pelerin Society (MPS)**[S], *founded by*[P] **Friedrich v. Hayek in 1947**[O], [...]	Ausgehend von den Mont Pelerin Society (MPS), die 1947 von Friedrich v. Hayek gegründet wurde, [...]
In his view **they**[S] would finally *lead to*[P] ***'The Road to Serfdom'***[O], that is the title of his famous book published in 1944	**They**[S] *would lead*[P] to the ***'road of servitude'***[O] according to the title of his book published in 1944	Sie würden auf den 'Weg der Knechtschaft' führen, so der Titel seines 1944 veröffentlichten Buches

6.6 Political sciences

An example statement about the funding of the Mont Pelerin Society can be found in Table 11. Here, OpenIE6 yielded nearly the same statement for both versions, the English and German-to-English translation. The only difference was the detail, e.g., that Friedrich v. Hayek was an Austrian economist and philosopher, which was not included in the German text.

Another example about the *Road of Serfdom*, a famous book, revealed problems with the translation. The German word *Knechtschaft* was translated into *servitude*, which was not the correct title of the book (*Serfdom*). However, the extraction that *they lead to the road of Serfdom/servitude* was similar. OpenIE6 although extracted correctly that the *famous book* or *his book* was published in *1944* for both versions.

Another long English sentence was: *This article deals with the role of policy learning for the genesis of Austrian art policy during the 1980ies and early 1990ies and seeks to utilize the conclusion drawn from this analysis for the further development of the concept of policy learning.* The German version *Dieser Artikel befasst sich mit der Rolle des Policy Learning für die Genese der österreichischen Kunstpolitik in den 1980er und frühen 1990er Jahren und versucht, die Schlussfolgerungen aus dieser Analyse für die Weiterentwicklung des Konzepts des Policy Learning zu nutzen.* was translated in *This article addresses the role of policy learning in the formation of Austrian reproductive technology policy during the 1980s and early 1990s and seeks to make findings in this regard useful for a further development of the*

conception of policy learning. OpenIE6 then extracted four extractions for each sentence, respectively. Three of these statements were nearly identical except for some wording. The fourth statement differed in the level of detail in the object phrase: *This article seeks* (*to utilize the conclusion* vs. *to utilize the conclusion drawn from this analysis for the further development of the concept of policy learning*).

7 Discussion

In the following, we discuss how suitable nearly unsupervised extraction workflows are in digital libraries by considering technical and conceptual limitations. Furthermore, we give best practices on what to do and when supervision is necessary.

7.1 Toolbox improvements

The toolbox filtered verb phrases by removing non-verbs (stop words, adverbs, etc.) and verbs like *be* and *have*. Here negations in verb phrases were lost, too. We implemented a parameter to make this behavior optional. Next, we implemented the subject entity filter that was useful in Wikipedia and Political Sciences. Here a statement's subject must be linked to an entity, but the object can keep the original information. In particular, when subject noun phrases were short and object noun phrases were complex, the subject filter could be used to construct a semi-structured knowledge base, e.g., showing all actions of *Albert Einstein* or *positions* that the *EU* has taken. In addition, we implemented a clustering-based canonicalization procedure like proposed by [31].

7.2 Technical toolbox limitations

In addition, the dictionary-based entity linker fails to resolve short and ambiguous mentions. These wrongly linked mentions cause problems in the cleaning step (entity-based filters). Here, more advanced linkers would be more appropriate to improve the overall quality. A co-reference resolution is also missing, i.e., resolving all pronouns and mentions that refer to known entities. PathIE is currently restricted to binary relations but might be extended to extract more higher-ary relations, e.g., by considering all connected entities via a verb phrase or a particular keyword like treatment. A suitable cleaning would be possible if the relation arguments (subject and object) could be restricted to entity types.

7.3 Restrictions of unsupervised IE

The first significant restriction of unsupervised methods is their focus on and thus restriction to grammatical structures. Suppose the example: *The German book Känguru-Chroniken was written by Marc-Uwe Kling*. Here unsupervised methods may not extract that the language of the work is German.

In common relation extraction benchmarks, such relations appear and can be learned and inferred by modern language models [5, 21]. However, we argue that such extractions require high domain knowledge, typically unavailable in unsupervised extraction methods. Similar examples could be made in specialized domains like Pharmacy (treatments, inhibitions, etc.). Moreover, it is not possible to integrate this knowledge into unsupervised models by design: The model would need training data to infer such rules and, thus, be supervised. We do not expect unsupervised models with access to comprehensive domain-specific knowledge soon. And even if applying such a model in a new domain with new types of relations would then again require a re-training of that model, e.g., for treatment relations in Pharmacy.

Our case studies showed that OpenIE6 extracts noun phrases in two ways: Either noun phrases are short and miss relevant information from the sentence. These phrases are easier to handle but may be unhelpful in the end. Or the noun phrases are long and complex but retain the original information. Indeed, our analysis in Sect. 4.4 revealed that many noun phrases, especially objects, were complex. Handling complex phrases requires more advanced cleaning methods. Although CoreNLP OpenIE extracted less complex noun phrases, the overall problem of how to handle such noun phrases still remained.

The toolbox canonicalization procedure for relations considers only the verb phrases, not the surrounding context. Verb phrases like *uses*, *publish*, and *prevent* could refer to a plethora of relations. In the end, more advanced methods are required for a suitable canonicalization quality. Even clustering-based methods will not solve this issue by design, if the sentence context is not considered. Especially, canonicalizing OpenIE6 verb phrases to precise relations was not really possible.

7.4 Handling non-english texts

Although our case study in Sect. 6 was preliminary, it showed the potential of modern machine translation. Even complicated and nested sentences were well translated, and the information extraction method yielded similar extractions in all three domains. Instead of acquiring cost-intensive training data to train information extraction models for non-English languages, translating such languages to English could be a suitable alternative here. However, performing translations could still be challenging if languages are underrepresented.

7.5 Application and costs

Although we observed several issues and limitations, these methods can be used to implement services in digital

Table 12 The table summarizes the measured runtimes for the samples and gives an estimation for the whole collection

		Wikipedia		Pharmacy		Political sciences	
		Sample	Estimation	Sample	Estimation	Sample	Estimation
2013 Server – Nvidia GTX 1080 TI & 2xCPU (8/16) & 377GB DDR3 Memory							
Entity Det.	*NER*	10.5 min	19.4 days	–	–	10.1 min	21.6 h
	EL	0.6 min	1.2 days	1.2 min	2.8 days	0.7 min	1.4 h
Extraction	*PathIE*	2.6 min	4.7 days	2.0 min	4.6 days	–	–
	OpenIE6	53.6 min	98.8 days	74.0 min	170.0 days[14]	55.4 min	5.0 days
	CoreNLP	6.7 min	12.2 days	7.3 min	16.6 days	5.0 min	11 h
Cleaning		< 1 h	< 1 day	< 1 h	< 1 day	< 1 h	< 1 day
2021 Server – Nvidia A40 & 2xCPU (24/48) & 2TB DDR4 Memory							
Entity Det.	*NER*	4.0 min	7.4 days	–	–	3.8 min	8.2 h
	EL	4.2 sec	3.1 h	9.0 sec	8.3 h	5.9 sec	14.4 min
Extraction	*PathIE-32*	1.5 min	2.7 days	1.2 min	2.9 days	–	–
	PathIE-96	2.3 min	4.3 days	2.6 min	5.9 days	–	–
	OpenIE6	18.6 min	34.3 days	26.2 min	60.1 days	19.6 min	1.8 days
	CoreNLP	3.3 min	6.1 days	3.3 min	7.5 days	2.3 min	5.0 h
Cleaning		< 1 h	< 1 day	< 1 h	< 1 day	< 1 h	< 1 day

libraries. We summarize the measured runtimes and computed estimations for the corresponding collections in Table 12.

Consider our PubPharm project, for example: PathIE could enable a graph-based retrieval service with moderate costs [16]. Around nine sessions with experts and moderate development time were necessary to implement a workflow. The computation of PathIE took 2 min on our sample and was estimated to take 4.6 days for the whole PubMed collection. Indeed, PubPharm could perform the complete extraction workflow in one week.

Our current cooperation with Pollux revealed that OpenIE6 could bring more structure to this domain. We will continue our work with Pollux by focusing on research questions that we would like to answer with semi-structured information derived from OpenIE6 with subject filtering.

On our server with an Nvidia GTX 1080 TI, the computation of OpenIE6 took 55.4 min on the Pollux sample and is estimated to take five days for the complete collection. For Wikipedia the sample took 53.6 min, and all English articles would require 98.8 days. Note that we used a single GPU from 2016. Hence the workflow can be accelerated with a modern GPU and parallelized by utilizing multiple GPUs. In addition, OpenIE6 can also be restricted to sentences that contain at least two entities. Here the runtime was decreased from 55.4 to 22.4 min (Pollux) and 53.6 to 41.4 min (Wikipedia). CoreNLP OpenIE took way less time than OpenIE6, i.e., was estimated to take 12.2 days for the complete Wikipedia, 16.6 days for the PubMed corpus, and 11 h for the Political Sciences corpus.

7.5.1 Server 2021

As an extension, we measured the runtime performance on our latest server from 2021. In contrast to our old server, this had two Intel(R) Xeon(R) Gold 6336Y CPU @ 2.40GHz (24 cores and 48 threads each), 2TB DDR4 main memory, and nine Nvidia A40 GPUs with 48GB memory. Note that we only utilized a single GPU for this comparison. Again, the runtimes are reported in Table 12. The main finding here was that the runtime was decreased in GPU-intensive tasks (NER or OpenIE6) by a factor of about three. In CPU-intensive tasks (EL + PathIE + CoreNLP OpenIE), we utilized all CPU threads (96). The entity linking runtime decreased clearly and was estimated to take less than a half day for all three domains. CoreNLP OpenIE achieved a speedup of about a factor of two. And for PathIE, we made an unexpected observation: Utilizing all 96 threads took about double the time than utilizing only 32 threads. PathIE utilizes the Java Stanford CoreNLP tool for generating sentence dependency parses, which might not scale well or might have resource-limited boundaries (e.g., I/O from disk).

7.6 Best practices

Subsequently, we give some advice that we can deduce from our case studies. OpenIE6 handles short and simple sentences well. Here the exact entity filter will produce suitable extrac-

[14] We wrongly reported 98.8 days in [17].

tions but decrease the recall drastically. The partial entity filter improves the recall but often messes up the original information. We recommend two strategies for long and complex sentences:

First, do not use the exact or partial filter because important information can be missed. Use the subject filter to retrieve precise entities as subjects and the original information in objects. This filter allows the construction of semi-structured knowledge bases, e.g., positions that were taken by the *EU* or actions that *Albert Einstein* has done. Another option is to use no filter, but then, the extractions are not cleaned in any way.

Second, PathIE can find specialized relations that are expressed by keywords, e.g., treatment and therapy. But PathIE requires directed relations that must be cleaned by entity type constraints. Detecting such relations via PathIE is fast and probably cheaper than training supervised extraction models. However, PathIE will fail if several entities of the same type are mentioned within a sentence, e.g., side effects of treatments. Here supervised methods are required to achieve suitable quality. Another limitation of PathIE and our canonicalization procedures is that a verb phrase/keyword must refer to a single relation. A verb phrase like *use* that refers to a plethora of different relations could, in this way, hardly be canonicalized, regardless of whether we used a relation vocabulary-based or a clustering-based approach. For such cases, the context of the sentence, and thus, supervision is necessary to extract the underlying relation reliably.

8 Conclusion

In this paper, we have studied nearly unsupervised extraction workflows for a practical application in digital libraries. We focused on three different domains to generalize our findings, namely the encyclopedia Wikipedia, Pharmacy, and Political Sciences. First, the scalability of the investigated methods was acceptable for our partners. Second, unsupervised extraction workflows required intensive cleaning and canonicalization to result in precise semantics. Thus they do not work out-of-the-box, and reliably canonicalizing OpenIE verb phrases remains an open issue because contexts are not considered by relation vocabulary/clustering methods. Although such cleaning can be exhausting, the pharmaceutical case study yielded a novel retrieval service. Such a service would not have been possible when training data must have been collected for each relation. In addition, not filtering complex object phrases can allow the construction of semi-structured knowledge bases or enrich the original texts, e.g., show all actions of Albert Einstein. In conclusion, unsupervised extraction workflows are worth studying in digital libraries, even if, the library contains non-English texts. Those workflows come with limitations and require cleaning, but they entirely bypass the lack of training data in the extraction phase.

Supplementary information

The code of the extraction toolbox and the case study can be found in our GitHub repository.[15] An archived version can be found in the Software Heritage.[16]

Acknowledgements Supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): PubPharm - Specialized Information Service for Pharmacy (Gepris 267140244). We would also like to thank Pollux—Specialized Information Service for Political Science for providing the data for our case study, and Wolfgang Otto (GESIS) for supporting our evaluation.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Attardi, G.: Wikiextractor. https://github.com/attardi/wikiextractor (2015)
2. Auer, S., Bizer, C., Kobilarov, G., et al.: Dbpedia: A nucleus for a web of open data. In: The Semantic Web, pp 722–735. Springer Berlin Heidelberg, (2007). https://doi.org/10.1007/978-3-540-76298-0_52
3. Bhardwaj, S., Aggarwal, S., Mausam, M.: CaRB: A crowdsourced benchmark for open IE. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp 6262–6267. (2019).https://doi.org/10.18653/v1/D19-1651
4. Blasi, D., Anastasopoulos, A., Neubig, G.: Systematic inequalities in language technology performance across the world's languages. In: Proceedings of the 60th Annual Meeting of the ACL, pp 5486–5505. (2022). https://doi.org/10.18653/v1/2022.acl-long.376
5. Devlin, J., Chang, M.W., Lee, K., et al.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of

[15] https://github.com/HermannKroll/KGExtractionToolbox.

[16] https://archive.softwareheritage.org/swh:1:dir:5b575ac043e2bd61999250564a16a220c88ee5c9.

the ACL: Human Language Technologies, vol. 1, pp. 4171–4186 (2019). https://doi.org/10.18653/v1/N19-1423 https://doi.org/10.18653/v1/N19-1423
6. Gashteovski, K., Yu, M., Kotnis, B., et al.: BenchIE: A framework for multi-faceted fact-based open information extraction evaluation. In: Proceedings of the 60th Annual Meeting of the ACL, pp 4472–4490, (2022). https://doi.org/10.18653/v1/2022.acl-long.307
7. Groth, P., Lauruhn, M., Scerri, A., et al.: Open information extraction on scientific text: An evaluation. In: Proceedings of the 27th International Conference on Computational Linguistics, pp 3414–3423, (2018). https://aclanthology.org/C18-1289
8. Hristovski, D., Kastrin, A., Dinevski, D., et al.: Constructing a graph database for semantic literature-based discovery. Stud. Health Technol. Inform. **216**, 1094 (2015)
9. Jaradeh, M.Y., Oelen, A., Farfar, K.E., et al.: Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In: Proceedings of the 10th International Conference on Knowledge Capture. ACM, K-CAP '19, pp. 243-246. (2019). https://doi.org/10.1145/3360901.3364435
10. Kilicoglu, H., Shin, D., Fiszman, M., et al.: SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics **28**(23), 3158–3160 (2012). https://doi.org/10.1093/bioinformatics/bts591
11. Kolluru, K., Adlakha, V., Aggarwal, S., et al.: Openie6: iterative grid labeling and coordination analysis for open information extraction. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, pp. 3748–3761. (2020). https://doi.org/10.18653/v1/2020.emnlp-main.306
12. Kolluru, K., Mohammed, M., Mittal, S., et al.: Alignment-augmented consistent translation for multilingual open information extraction. In: Proceedings of the 60th Annual Meeting of the ACL, pp 2502–2517. (2022). https://doi.org/10.18653/v1/2022.acl-long.179
13. Kroll, H., Kalo, J.C., Nagel, D., et al.: Context-compatible information fusion for scientific knowledge graphs. In: Digital Libraries for Open Knowledge. Springer International Publishing, pp. 33–47. (2020). https://doi.org/10.1007/978-3-030-54956-5_3
14. Kroll, H., Al-Chaar, J., Balke, W.: Open information extraction in digital libraries: Current challenges and open research questions. In: Proceedings of the Workshop on Digital Infrastructures for Scholarly Content Objects (DISCO) co-located JCDL 2021, CEUR Workshop Proceedings, vol. 2976. CEUR-WS.org, pp. 14–18. (2021a). http://ceur-ws.org/Vol-2976/short-1.pdf
15. Kroll, H., Pirklbauer, J., Balke, W.: A toolbox for the nearly-unsupervised construction of digital library knowledge graphs. In: ACM/IEEE Joint Conference on Digital Libraries, JCDL 2021. IEEE, pp. 21–30. (2021b). https://doi.org/10.1109/JCDL52503.2021.00014
16. Kroll, H., Pirklbauer, J., Kalo, J., et al.: Narrative query graphs for entity-interaction-aware document retrieval. In: Towards Open and Trustworthy Digital Societies - 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021., Lecture Notes in Computer Science, Vol 13133. Springer, pp. 80–95. (2021c). https://doi.org/10.1007/978-3-030-91669-5_7
17. Kroll, H., Pirklbauer, J., Plötzky, F., et al.: A library perspective on nearly-unsupervised information extraction workflows in digital libraries. In: Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries. ACM, JCDL '22, (2022a). https://doi.org/10.1145/3529372.3530924
18. Kroll, H., Plötzky, F., Pirklbauer, J., et al.: What a publication tells you-benefits of narrative information access in digital libraries. In: Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries. ACM, JCDL '22, (2022b). https://doi.org/10.1145/3529372.3530928
19. Kroll, H., Pirklbauer, J., Kalo, J.C., et al.: A discovery system for narrative query graphs: entity-interaction-aware document retrieval. Int. J. Digit. Libr. (2023). https://doi.org/10.1007/s00799-023-00356-3
20. Kruiper, R., Vincent, J., Chen-Burger, J., et al.: In layman's terms: semi-open relation extraction from scientific texts. In: Proceedings of the 58th Annual Meeting of the ACL, pp. 1489–1500. (2020). https://doi.org/10.18653/v1/2020.acl-main.137
21. Lee, J., Yoon, W., Kim, S., et al.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2019). https://doi.org/10.1093/bioinformatics/btz682
22. Liu, Y., Bai, K., Mitra, P., et al.: Tableseer: automatic table metadata extraction and searching in digital libraries. In: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries. ACM, JCDL '07, p 91-100, (2007). https://doi.org/10.1145/1255175.1255193
23. Manning, C.D., Surdeanu, M., Bauer, J., et al.: The stanford corenlp natural language processing toolkit. In: Proceedings of the 52nd Annual Meeting of the ACL, ACL 2014. ACL, pp 55–60, (2014). https://doi.org/10.3115/v1/p14-5010
24. Mendez, D., Gaulton, A., Bento, A.P., et al.: ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res. **47**(D1), D930–D940 (2018). https://doi.org/10.1093/nar/gky1075
25. Mikolov, T., Chen, K., Corrado, G., et al.: Efficient estimation of word representations in vector space. In: 1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings, (2013). http://arxiv.org/abs/1301.3781
26. Niklaus, C., Cetto, M., Freitas, A., et al.: A survey on open information extraction. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 3866–3878. (2018). https://aclanthology.org/C18-1326
27. Qi, P., Zhang, Y., Zhang, Y., et al.: Stanza: A python natural language processing toolkit for many human languages. In: Proceedings of the 58th Annual Meeting of the ACL: System Demonstrations, pp. 101–108. (2020). https://doi.org/10.18653/v1/2020.acl-demos.14
28. Sai, STYS., Chakraborty, P., Dutta, S., et al.: Joint entity and relation extraction from scientific documents: Role of linguistic information and entity types. In: Proceedings of the 2nd Workshop on EEKE co-located with JCDL 2021, CEUR Workshop Proceedings, vol 3004. CEUR-WS.org, pp. 15–19. (2021). http://ceur-ws.org/Vol-3004/paper2.pdf
29. Schardelmann, T., Otto, W.: Pollux - von der bedarfsanalyse zur technischen umsetzung. Bibliotheksdienst **52**(3–4), 225–234 (2018). https://doi.org/10.1515/bd-2018-0029
30. Thilakaratne, M., Falkner, K., Atapattu, T.: Information Extraction in Digital Libraries: First Steps towards Portability of LBD Workflow, ACM, pp. 345-348. (2020). https://doi.org/10.1145/3383583.3398607
31. Vashishth, S., Jain, P., Talukdar, P.: Cesi: Canonicalizing open knowledge bases using embeddings and side information. In: Proceedings of the 2018 World Wide Web Conference. WWW S. Committee, WWW '18, pp. 1317-1327. (2018). https://doi.org/10.1145/3178876.3186030
32. Vrandecic, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014). https://doi.org/10.1145/2629489
33. Wei, C., Kao, H., Lu, Z.: Pubtator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res. **41**(W1), 518–522 (2013). https://doi.org/10.1093/nar/gkt441
34. Wei, C., Allot, A., Leaman, R., et al.: Pubtator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res. 47(W1):W587–W593. (2019). https://doi.org/10.1093/nar/gkz389

35. Weikum, G., Dong, X.L., Razniewski, S., et al.: Machine knowledge: creation and curation of comprehensive knowledge bases. Foundations and Trends in Databases (2021). https://doi.org/10.1561/1900000064
36. Williams, K., Wu, J., Wu, Z., et al.: Information extraction for scholarly digital libraries. In: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries. ACM, JCDL '16, pp. 287-288. (2016). https://doi.org/10.1145/2910896.2925430
37. Zhang, R., Cairelli, M.J., Fiszman, M., et al.: Using semantic predications to uncover drug-drug interactions in clinical data. J. Biomed. Inform. **49**, 134–147 (2014). https://doi.org/10.1016/j.jbi.2014.01.004
38. Zhang, Y., Chen, Q., Yang, Z., et al.: Biowordvec, improving biomedical word embeddings with subword information and mesh. Sci. Data **6**(1), 1–9 (2019)