Exploring the Role of Generative AI in Constructing Knowledge Graphs for Drug Indications with Medical Context

Reham Alharbi[1,*,†], Umair Ahmed[2,†], Daniil Dobriy[3,†], Weronika Łajewska[4,†], Laura Menotti[5,†], Mohammad Javad Saeedizade[6,†] and Michel Dumontier[7,†]

[1]*Department of Computer Science, University of Liverpool, Liverpool, UK*
[2]*School of Science and Technology, University of Camerino, Camerino, Italy*
[3]*Institute for Data, Process and Knowledge Management, Vienna University of Economics and Business, Vienna, Austria*
[4]*Department of Electrical Engineering and Computer Science, University of Stavanger, Stavanger, Norway*
[5]*Department of Information Engineering, University of Padua, Padova, Italy*
[6]*Department of Computer Science, Linköping University, Linköping, Sweden*
[7]*Institute of Data Science, Maastricht University, Maastricht, The Netherlands*

Abstract

The medical context for a drug indication provides crucial information on how the drug can be used in practice. However, the extraction of medical context from drug indications remains poorly explored, as most research concentrates on the recognition of medications and associated diseases. Indeed, most databases cataloging drug indications do not contain their medical context in a machine-readable format. This paper proposes the use of a large language model for constructing DIAMOND-KG, a knowledge graph of drug indications and their medical context. The study 1) examines the change in accuracy and precision in providing additional instruction to the language model, 2) estimates the prevalence of medical context in drug indications, and 3) assesses the quality of DIAMOND-KG against NeuroDKG, a small manually curated knowledge graph. The results reveal that more elaborated prompts improve the quality of extraction of medical context; 71% of indications had at least one medical context; 63.52% of extracted medical contexts correspond to those identified in NeuroDKG. This paper demonstrates the utility of using large language models for specialized knowledge extraction, with a particular focus on extracting drug indications and their medical context. We provide DIAMOND-KG as a FAIR RDF graph supported with an ontology. Openly accessible, DIAMOND-KG may be useful for downstream tasks such as semantic query answering, recommendation engines, and drug repositioning research.

Keywords
Knowledge Graph Construction, LLMs in KGC, Medical Knowledge Graph

---

CEUR Workshop Proceedings (CEUR-WS.org)

1. Introduction

Drug indications are regulatory-approved uses for a medicine. The indication of a drug, along with its medical context better defines its therapeutic intent and provides additional prescribing guidance for physicians. Valid medical contexts include the underlying medical illness for the *target condition* ("**Quetiapine** tablets are indicated for the **acute treatment of manic episodes associated with bipolar I disorder**"), the *age group* ("**EMGALITY** is indicated for the treatment of episodic cluster headache in **adults**"), and *co-therapies*, i.e. drugs that should be administered in combination, ("**Clonidine hydrochloride** injection is indicated **in combination with opiates** for the treatment of severe pain in cancer patients"). The drug indication and its medical context are contained within drug product labels whose contents are subject to approval by regulatory agencies such as the U.S. Food and Drug Administration (FDA) or the European Medicines Agency (EMA).

Machine-readable representations of drug indications are key in computational drug discovery [1] and clinical decision-making [2]. However, while databases have been created to store drug indications in a machine-readable manner (e.g. DrugCentral [1]), these either do not contain their medical context and as such do not correctly cover the therapeutic intent, or the medical context is available in some form of natural language and is not available for computation. The lack of a computable medical context will necessarily limit the development of accurate methods to predict new drug uses [2] or make treatment recommendations. As a case example, the Oncology Expert Advisor system derived from IBM Watson was pulled from the market after making unsafe suggestions relating to cancer treatment.[1] Thus, medical context should be taken into account while creating accurate, precise, and compute-accessible representations of drug indications. Knowledge Graphs (KGs) serve as a natural and intuitive way to store, query, and explore structured knowledge. However, building high-quality knowledge graphs requires substantial manual effort. Thus, automated Knowledge Graph Construction (KGC) methods have emerged to alleviate the burden of manual data curation [3]. To this end, Large Language Models (LLMs) are promising technologies for natural language understanding, and in particular, for constructing knowledge graphs. Recent unpublished work suggests that carefully engineered prompts have the potential to extract entities and their relations in a structured manner [4, 5]. To the best of our knowledge, there is no method that correctly extracts the medical context for drug indications.

Contributions In this work, we propose a novel approach to use a LLM that extracts drugs, their indications, and the associated medical context to a target Knowledge Graph. We aim to answer the following research questions: *"RQ1: To what extent does the addition of more instruction to the LLM yield a more accurate/complete extraction of the context?"*, *"RQ2: How many drug indications include a medical context?"* and *"RQ3: What is the quality of the generated knowledge graph?"*. The main contributions of the research are as follows: (i) the development of a LLM-based framework to extract triples and their context, perform entity recognition, and produce a valid RDF graph (ii) the use of the framework to extract drug indications and their medical context from sentences in natural language (iii) an evaluation of prompts that vary in

---

[1]https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/

their specificity (iv) an estimate of the scale of the problem of recognising and representing medical context

The rest of the paper is structured as follows: Section 2 reports previous efforts in this task and useful resources. Section 3 defines the proposed framework and all its components. Section 4 analyzes the results and provides an empirical evaluation of our method. Finally, Section 5 concludes the paper with some final remarks.

2. Related work

DailyMed[2] is a repository of drug labels approved by the FDA and hosted by the National Institute of Health (NIH). It contains a wide range of information about drug indications and contradictions, and is available as XML files available through an API. The task of synthesizing information about the therapeutic intent has been addressed in several works. Drugbank [6] is a curated resource for detailed drug (i.e. chemical) data along with their drug targets (i.e. protein), but the indication, if present, is solely expressed in natural language. DrugCentral [1] provides drug information along with indications extracted from product labels, however, not all indications are coded and there is no additional medical context available. Névéol and Lu [7] focus on automatically extracting and integrating drug indication information from multiple health resources such as DailyMed and MeSH Scope notes. MEDI [8] applies NLP and ontology relationships to extract indications for single-ingredient medications. Prompt-based methods using LLMs are being explored for clinical concept extraction [9] and Drug-Drug interaction triplets extraction [10]. To the best of our knowledge, none of the automated methods extracting drug indications take into account the medical context.

Several efforts have been directed towards semi-automated curation of drug indications. LabeledIn [11] is a human-reviewed, machine-readable, and source-linked catalogue of 7805 indications for 250 human drugs. However, it does not curate medical context. InContext [12] is a curated set of indications and their medical context for 150 drugs. The dataset was curated by annotating Dailymed HTML pages using Hypothesis.is [13] and the National Center for Biomedical Ontologies (NCBO) BioPortal annotator [14] for concept recognition. InContext defines five medical contexts: (i) *Co-prescribed medication* - drugs commonly prescribed together; (ii) *Co-therapies* - procedures or therapies (not drug-related, i.e. radiotherapy) that should be applied in combination with the drug; (iii) *Co-morbidities* - diseases or conditions that commonly occur together (with a target condition) in the same patients; (iv) *Genetics* - genetic variants for a given disease; (v) *Temporal aspects* - information that explains at what life stage, disease stage, or treatment phase a drug should be administered. NeuroDKG [15] is a Knowledge Graph containing drug indications and their medical context for 101 drugs (from a total of 174 sentences) that target neurological disorders.

3. Proposed approach

We propose a framework to transform a textual description of drug indications with medical context into a knowledge graph using LLM-powered entity recognition along with identifier

---

[2]https://dailymed.nlm.nih.gov/dailymed/
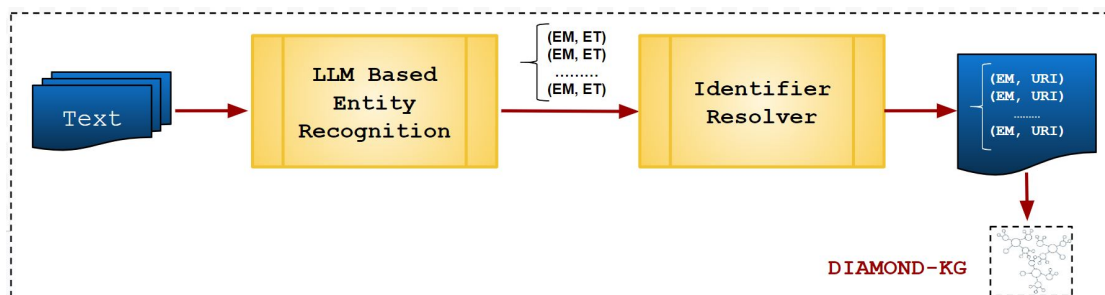
Figure 1: The DIAMOND-KG framework, which includes LLM-based entity recognition, the Identifier Resolver, and the final output presented as a KG.

resolution (see Figure 1).

3.1. LLM-based entity recognition

We designed a set of prompts to guide a LLM to identify and extract entities of interest. In particular, we focus on extracting contextual information from sentences to enrich drug-disease interaction triples. The prompts contain a *Paragraphs* section and a *Instructions* section. Each prompt is extended with additional specification to return a defined JSON-formatted object. The prompts' instructions range from general instructions (prompt 1) to specific (prompt 3). The first prompt, called ""`Triple Prompt`" extracts subject-verb-object triples from a given text. Subsequently we developed the "`Context Prompt`", which identifies the context from a given text and enriches the original triples with such information. Finally, the "`Medical Context Prompt`" identifies specific medical context types. This prompt contains a *Definition of context types* section within the *Instructions* section. All prompts and the complete list of predicates describing the medical context for drug indications are available in the DIAMOND-KG GitHub repository in the "Supplemental Material" directory[3].

3.2. Identifier Resolver

The values for each context type are then grounded to database/ontology identifiers using the NIH NCATS Translator SRI Name Resolution API [4]. The name resolution service takes lexical strings and attempts to map them to identifiers (CURIES; composed of a prefix for the source followed by a delimiter followed by the resource identifier) from a vocabulary or ontology. The lookup is not exact but includes partial matches. For each entity mention, we obtain a list of 5 results representing possible conceptual matches, of which the first is the preferred choice, and the remainder are ranked by the next preferred resource.

---

[3]https://github.com/semantisch/diamond-kg/tree/main/Supplementary%20Material
[4]https://name-resolution-sri.renci.org/docs

3.3. RDF Graph Generation

The last component of the framework is the generation of an RDF graph from the JSON result containing the extracted and named entities and relations. We define the DKG namespace[5] and iteratively process the sentences from the JSON results, constructing a graph that follows a lightweight Diamond-KG ontology described below to represent the sentences and their semantic components. For each sentence (a *dkg:sentence*) and each identified component of a sentence (a *dkg:part*), we mint a unique IRI in the DKG namespace, using a hashing algorithm for collision resistance, resuting in *dkg:[HASH]* and *dkg:part/[HASH]* minted IRIs respectively, and assign them corresponding label values via *rdfs:label*. Next, we relate the components to their sentence of origin via a *dkg:hasPart* relation. In the case of prompt 1, the components form asserted triples. For prompt 2 and prompt 3, we classify the components according to their context type. Namely, utilizing the IRIs *dkg:freecontext/[LABEL]* for the prompt 2 contexts and *dkg:definedcontext/[LABEL]* for the prompt 3 contexts respectively. Finally, the components are grounded to the database/ontology identifiers (from the NCATS Translator SRI Name Resolution service) via a *skos:closeMatch* relation. The resulting graph is made publicly available in Turtle format[6] and documented following the FAIR (Findable, Accessible, Interoperable, Reusable) Data Guidelines [16]. Specifically, each entity is identified using a unique Uniform Resource Identifier (URI) for each semantic component within the graph. To ensure data is both interoperable and reusable, the graph is enriched with comprehensive metadata that adheres to a set of standard vocabularies. The resulting DIAMOND-KG comprises 12,363 triples, 1,186 entities, 435 predicates, and 148 classes. While the underlying DIAMOND-KG ontology totals 15 classes (namely, *dkg:Sentence*, *dkg:Context*, *dkg:Free_context*, *dkg:Defined_context* as well as the 11 defined contexts) and 1 predicate (*dkg:hasPart*), the majority of DIAMOND-KG classes (133) stem from the free contexts generated in prompt 2 and the majority of predicates (430) from the asserted triples generated in prompt 1.

The implementation is in Python 3.6 and published as a GitHub repository[7] under the MIT license[8]. We use OpenAI *gpt-4* model API as the LLM. The number of maximum requested tokens (with 1 token approximately corresponding to 4 chars of English text[9]) is set to 6,144 (current maximum value for the *gpt-4* model) to allow batch prompts containing many paragraphs as input. The paragraphs are processed individually (1 paragraph per prompt) as batch processing for this prompt has been found to lead to complex interactions in the prompt results due to context categories being conceptualised for the batch of sentences together. We use RDFLib 6.3.2 to generate the RDF graph.

---

[5]Full IRI: http://purl.org/dkg/v1/

[6]https://huggingface.co/datasets/um-ids/diamond-kg

[7]Available at https://github.com/semantisch/diamond-kg

[8]Refer to https://opensource.org/licenses/MIT

[9]For detailed token calculation refer to: https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them

Table 1

Results statistics. For each prompt, we provide the total number of extracted triples directly related to context (column "Tot. Triples"), the average number of triples for each sentence (column "Avg. T/S"), and the percentage of sentences that contain context information (column "Context Freq."). The first row reports the same statistics for the NeuroDKG dataset. The symbol "—" represents data not available.

Prompt	Tot. Triples	Avg. T/S	Context Freq.
NeuroDKG	510	3.0	64.11%
Prompt 1	442	2.11	—
Prompt 2	911	4.82	—
Prompt 3	621	3.32	71.12%

4. Results and Discussion

4.1. Experimental Setup

We compare our results to the NeuroDKG dataset [15], which contains 174 sentences concerning indications of neurological drugs. The manually-annotated triples are subsequently used to build a KG about drug indications with medical context which comprises 2,397 triples, 460 entities, 13 properties, and 10 classes[10]. To provide a comparable ground truth, we restricted NeuroDKG to consider only triples of interest, i.e. those in common with DIAMOND-KG. This step resulted in a dataset of 510 triples, with an average of 3.0 triples for each sentence. The complete list of selected context-related predicates from NeuroDKG that can be mapped to the medical context extracted by the DIAMOND-KG prototype is available in our GitHub repository in the "Supplemental Material" directory[11]. In addition, we map NeuroDKG's "disease" predicate to "target" in DIAMOND-KG, since they both represent the medical condition that is targeted by the considered drug. As far as context is concerned, 64.11% of sentences from NeuroDKG contain at least one triple representing medical context.
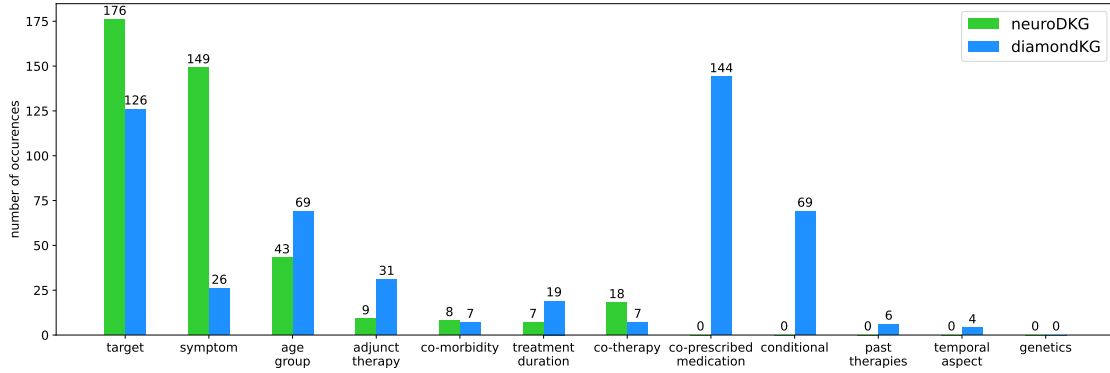
4.2. Prompts Analysis

For each prompt, Table 1 shows the number of triples relating to context that can be successfully extracted from sentences, the average number of triples per sentence, and the percentage of sentences that contain medical context information. The latter column is applicable only to the third prompt, as we provide a predefined set of predicates to extract. We also report the same information for the NeuroDKG dataset as a comparison. For comparability, we only include triples directly relating to context and do not include those triples that arise from our modelling choices in the construction of DIAMOND-KG (i.e., triples relating semantic components to sentences, labelling, entity typing and *rdfs:subClassOf* assertions).

The first prompt generates triples from a given text in a JSON format, without providing any domain-specific information. With this approach, we are able to extract 442 triples, with an average of 2.11 triples per sentence which is the lowest among the different prompts. This result can be attributed to the general nature of the instruction which may affect the system's

---

[10]The NeuroDKG Knowledge Base is available in zenodo at https://doi.org/10.5281/zenodo.5541440

[11]https://github.com/semantisch/diamond-kg/tree/main/Supplementary%20Material

Figure 2: Context information distribution from prompt 3 results. For each context entity, we report the number of sentences for which such information has been extracted by the prompt.

performance. The second prompt tests the power of generative AI to extract context information from text, without providing any specific types of context. As reported in Table 1, prompt 2 is able to extract 911 triples, with an average of 4.82 triples per sentence. The LLM-based Entity Recognition module identifies 140 different context types ranging from domain-specific entities such as "medication", "treatment", and "symptom" to more general entities like "demographics" or "Publication". We discover that most predicates are synonyms, e.g. "treatment" and "Medical treatment", or they are written in different ways, like "Medical Treatment" and "Medical_treatment" are considered two distinct contexts. This situation can lead to inconsistency among the KG's predicates and increases the possibility of duplicate information. To mitigate this issue, additional post-processing steps will be needed between the output of the LLM-based Entity Recognition module and the construction of the KG. The third prompt includes domain-specific information and the set of context entities we plan to extract. With this approach, we are able to extract 621 triples, with an average of 3.32 extracted triples for each sentence. Limiting the list of possible predicates results in a decrease in the number of extracted triples, which may also lead to lost information. Section 4.3 shows that we only lose little information with this approach. The third prompt identifies 71.12% sentences containing medical context, which means that we are able to extract context information for more sentences than the NeuroDKG dataset.

Figure 2 reports the distribution of the different predicates across the dataset, compared to the NeuroDKG output. Due to its nature, the NeuroDKG dataset extracts context information for the first seven context types. Overall, DIAMOND-KG extracts more context information than NeuroDKG in most cases. We have a similar number of context information extracted for types **"co-morbidity"** and **"co-therapy"**. The main difference between the two refers to the **"target"** and **"symptom"** context types. In the former case, NeuroDKG extracts such information for 176 sentences, compared to the 126 identified by DIAMOND-KG. This behaviour can be attributed to the variety of context types available to the DIAMOND-KG system. To this end, we analyzed the difference in values for the **"target"** context type and discovered that some triples labelled as "target" in NeuroDKG are identified by DIAMOND-KG with some other context types, such as **"symptom"** or **"co-morbidity"**. Take the sentence "*Xyrem is indicated for the treatment of*

cataplexy or excessive daytime sleepiness (EDS) in patients 7 years of age and older with narcolepsy"
as an example. NeuroDKG identifies as target "narcolepsy" and as symptom "cataplexy". On
the other hand, DIAMOND-KG correctly classifies "cataplexy" as a symptom but "narcolepsy"
is recognized as a co-morbidity. This classification is not entirely incorrect, indeed "narcolepsy"
is not the disease targeted by the drug, which treats "cataplexy". In general, DIAMOND-KG
assigns different context types for 120 medical contexts compared to NeuroDKG. Additionally,
given the same type assigned, in 82 cases the identified medical context by DIAMOND-KG is
broader and more informative than the one provided in NeuroDKG.

Most sentences were associated with the following context entities: **'co-prescribed medica-
tion'** (144 sentences), **'target'** (126 sentences), **'conditional'** (69 sentences) and **'age group'** (69
sentences). The least amount of sentences were associated with **'past therapies'**, **'temporal
aspects'**, and **'genetics'**, which are present in 6, 4, and 0 sentences respectively. We analyzed
the sentences and confirmed that this trend represents the real distribution of the dataset.
Indeed, fewer indications present information related to past therapies or at what life stage or
disease stage a drug should be administered (i.e. temporal aspects). About genetics, we found
no information related to such context type in any sentence. This could be due to the nature of
the dataset, i.e. neurological drugs.

4.3. Emprical Evaluation of the Third Prompt

As we discussed above, NeuroDKG is not sufficient to evaluate the results of DIAMOND-KG as it
contains fewer context types, and in most cases, our system seems to provide more informative
triples than NeuroDKG. To evaluate the quality of the information extracted by prompt 3, we
manually annotated all sentences in NeuroDKG considering all context types in DIAMOND-KG,
and compared our ground truth with the output of our system. For each (context value, context
type) pair, we are interested in whether our system is able to extract meaningful information
and classify them with the correct context type. Overall, DIAMOND-KG achieved an accuracy
of 63.52% throughout all context types, with a hamming loss of 4.20%. We identified three
common errors: "wrong pairs" (18.24%) are those that are not present in the ground truth,
"misclassified pairs" (9.77%) are those present in the ground truth but with a different context
type, and "missing pairs" (8.47%) are pairs that are present in the ground truth but not in the
DIAMOND-KG's output. Table 2 reports the performance metrics for each context type, which
varies in terms of their value as well as their occurrence. Context types with scores about
70% also exhibit reasonable support, e.g. "**target**", "**age group**", and "**symptom**". The method
performs well on identifying the "**age group**" suited for a given drug, with precision, recall,
and F1-score above 90%. The lowest results are registered on the context types with the lowest
support, where few wrong pairs have a higher impact on performance. These findings may
indicate that some context types are underrepresented in the dataset. Recall is above 70% in
most cases, except for "**temporal aspect**" (67%), "**co-morbidity**" (28%), and "**past therapies**"
(17%) which all exhibit low support. A high recall confirms that the method is able to return
most of the relevant pairs, meaning that providing a predefined set of context types does not
hinder the system or cause loss of information.

Table 2

Third prompt evaluation. We report the Precision (Prec), Recall (Rec), and F1-score for each context type and for the average results. "Support" (Sup) represents the number of pairs considered. We omitted the "genetics" context type, as it has support equal to zero.

(a) Precision, Recall, and for each context type.

	Prec	Rec	F1-Score	Sup
Target	**0.81**	0.79	0.80	214
Symptom	**0.67**	0.70	0.68	66
Age Group	**0.91**	**0.96**	**0.93**	71
Adj. Therapy	0.39	**1.00**	**0.84**	8
Co-morb.	0.58	0.28	0.38	25
Treat. Duration	0.59	0.76	0.67	17
Co-therapy	0.73	**1.00**	**0.84**	8
Co-presc. Med.	0.43	0.83	0.57	12
Conditional	0.56	0.77	0.65	65
Past Therapies	0.14	0.17	0.15	6
Temp. Aspects	0.40	0.67	0.50	3

(b) Precision, Recall, and for the average results.

	Prec	Rec	F1-Score	Sup
Micro Avg.	0.69	0.78	0.73	502
Macro Avg.	0.56	0.72	0.61	502
Weight. Avg.	**0.72**	**0.78**	**0.74**	502
Samples Avg.	0.64	0.64	0.64	502

5. Conclusion and Future Work

We explore a novel approach that leverages LLMs to extract relevant information and the associated medical context from drug indications. To the best of our knowledge, this is the first effort to extract such information by means of LLMs. The prototype system called DIAMOND-KG uses a LLM to recognize entities, which are subsequently passed to a service to perform identifier mapping, and the final step creates a FAIR RDF knowledge graph that complies. In relation to RQ1, we find that the refinement of the contexts produces higher quality outcomes to manually curated datasets. Moreover, it identifies a broader set of contexts and more informative results. While this framework offers a promising approach to automatically extract drug indications and their medical context, it also raises the possibility for this framework to accurately and systematically extract a wide variety of contextual information for other context-dependent settings. In relation to RQ2, based on the set of sentences annotated in NeuroDKG, we find that at least 71.12% of sentences have at least one context, which is greater than the 64.11% reported in the manually annotated NeuroDKG. This result indicates that a significant proportion of drug indications do contain a medical context, and the framework is able to identify these to a greater extent than manual curation. In relation to RQ3, the quality of the extraction varies based on the context type, but we mainly attribute such oscillation to the difference in the support. DIAMOND-KG achieved a high recall, demonstrating that the system extracts a high portion of relevant information and experiences little information loss.

Acknowledgments

References

[1] S. I. Avram, et al., Drugcentral 2021 supports drug discovery and repositioning, Nucleic Acids Research 49 (2020) D1160 – D1169.

[2] S. J. Nelson, et al., Formalizing drug indications on the road to therapeutic intent, JAMIA 24 (2017) 1169–1172.

[3] S. Marchesin, et al., Building a large gene expression-cancer knowledge base with limited human annotations, Database J. Biol. Databases Curation 2023 (2023).

[4] M. Trajanoska, et al., Enhancing knowledge graph construction using large language models, 2023. `arXiv:2305.04676`.

[5] J. H. Caufield, et al., Ontogpt, 2023. URL: https://monarch-initiative.github.io/ontogpt/.

[6] D. S. Wishart, et al., Drugbank: a comprehensive resource for in silico drug discovery and exploration, Nucleic Acids Research 34 (2005) D668 – D672.

[7] A. Névéol, Z. Lu, Automatic integration of drug indications from multiple health resources, in: Proc. of the 1st ACM international health informatics symposium, 2010, pp. 666–673.

[8] W. Q. Wei, et al., Development and evaluation of an ensemble resource linking medications to their indications, JAMIA 20 (2013) 954 – 961.

[9] C. Peng, et al., Clinical concept and relation extraction using prompt-based machine reading comprehension, JAMIA 30 (2023) 1486–1493.

[10] H. Hu, et al., A generative drug–drug interaction triplets extraction framework based on large language models, Proc. of the Association for Information Science and Technology 60 (2023) 980–982.

[11] R. Khare, et al., Labeledin: Cataloging labeled indications for human drugs, Journal of biomedical informatics 52 (2014) 448–56.

[12] K. Moodley, et al., InContext: curation of medical context for drug indications, Journal of Biomedical Semantics 12 (2021) 2.

[13] Hypothesis, Hypothesis.is – Open Annotation Tool, 2023. URL: https://web.hypothes.is, accessed: 2023-06-13.

[14] BioPortal, Bioportal annotator, 2023. URL: https://bioportal.bioontology.org/annotator, accessed: 2023-06-13.

[15] J. Yang, et al., Publishing Medical Context of Neurological Drug Indications as a Knowledge Graph, Technical Report, Institute of Data Science, Maastrich University, Maastricht, the Netherlands, 2021. URL: https://github.com/MaastrichtU-IDS/neuro_dkg/blob/master/publication.pdf.

[16] M. D. Wilkinson, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (2016) 1–9.