Proceedings of the Federated Conference on Computer Science and Information Systems pp. 265–272
DOI: 10.15439/2019F3
ISSN 2300-5963 ACSIS, Vol. 18

Knowledge Extraction and Applications utilizing Context Data in Knowledge Graphs

Jens Dörpinghaus*, Andreas Stefan†
Fraunhofer Institute for Algorithms and Scientific Computing,
Schloss Birlinghoven, Sankt Augustin, Germany
Email: *jens.doerpinghaus@scai.fraunhofer.de, †andreas.stefan@scai.fraunhofer.de

***Abstract*—Context is widely considered for NLP and knowledge discovery since it highly influences the exact meaning of natural language. The scientific challenge is not only to extract such context data, but also to store this data for further NLP approaches. Here, we propose a multiple step knowledge graph-based approach to utilize context data for NLP and knowledge expression and extraction. We introduce the graph-theoretic foundation for a general context concept within semantic networks and show a proof-of-concept-based on biomedical literature and text mining. We discuss the impact of this novel approach on text analysis, various forms of text recognition and knowledge extraction and retrieval.**

CONTEXT is a widely discussed topic in text mining and knowledge extraction since it is highly relevant to mine the semantic correct sense of unstructured text. For example in [1], Nenkova and McKeown discuss the influence of context on text summarization. Ambiguity does not only appear for common language words, but especially in scientific context. The scientific challenge is not only to extract such context data, but also to store this data for further NLP approaches. Here, we propose a multiple step knowledge graph-based approach to utilize context data for NLP and knowledge expression. We present a proof of concept based on biomedical literature and show an outlook on further improvements towards next generation knowledge extraction for example for training approaches from artificial intelligence and machine learning.

Knowledge graphs play in general an important role in recent knowledge mining and discovery. A *knowledge graph* (sometimes also called a *semantic network*) is a systematic way to connect information and data to knowledge on a more abstract level than language graphs. It is thus a crucial concept on the way to generate knowledge and wisdom, to search within data, information and knowledge. The context is a significant topic to generate knowledge or even wisdom. Thus, connecting knowledge graphs with context is a crucial feature.

Here, we use a quite general definition of context data. We assume that every information entity can also be a context information for other entities. For example a document can also be a context for other documents (e.g. by citing or referring to the other publication). An author is both a meta information to a document, but also itself context (by other publications, affiliations, co-author networks, ...). Other data is more obvious a context: named entities, topic maps, keywords, etc. extracted with text mining from documents. But already relations extracted from a text may stand for themselves, occurring in multiple documents and still valuable without the original textual information.

Starting with a simple document graph, in a first step we add context meta information, see figure 1. This will lead to a first knowledge graph which can be used for a first context-based text mining approach. The text mining approach will add more context data, for example from ontologies or relations extracted from the text. The graph with the additional context data can be used as starting basis for more detailed text mining approaches utilizing the novel context data. This step can be redone several time.

In addition using a graph structure has several more advantages for knowledge extraction in biological and medical research. Here scientists are for example interested in exploring the mechanisms of living organisms and gaining a better understanding of underlying fundamental biological processes of life. Today the biomedical field mostly relies on systems biology approaches such as integrative knowledge graphs to decipher mechanism of a disease, by considering system as a whole which is considered as a holistic approach. In that, disease modeling and pathway databases play an important role. Knowledge Graphs built using Biological Expression Language (BEL, see www.openbel.org) is widely applied in biomedical domain to convert unstructured textual knowledge into a computable form. The BEL statements that form knowledge graphs are semantic triples that consist of concepts, functions and relationships [2]. In addition, several databases and ontologies implicitly form a Knowledge Graph. For example Gene Ontology, see [3] or DrugBank, see [4] or [5] cover a huge amount of relations and references to other fields.

Over the last few years new domain specific languages (DSL) and knowledge representations like BEL [6] have been proposed to publish and store this kind of statements and findings. There are still several crucial issues converting literature to knowledge. For example the quality and completeness of such networks has to be evaluated. And with this, to generate new knowledge the context of concepts in a Knowledge Graph has to be considered.

We will first of all give a preliminary overview about information theory and management. With this, we will introduce and discuss the novel approach of managing and mining

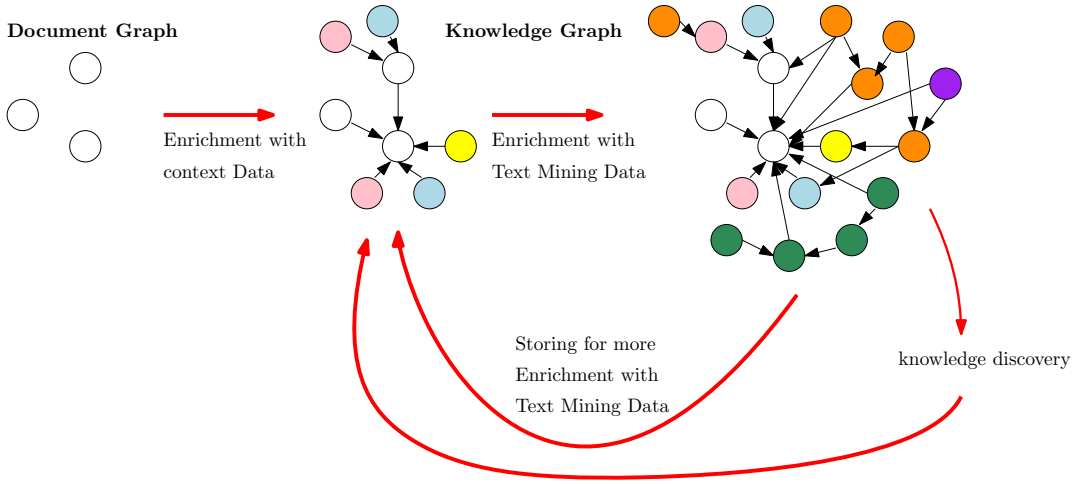

Fig. 1: Proposed workflow to extend a knowledge graph. First starting with a document graph, the basic meta information like authors, keywords etc. are added. This can be used as a basis for text mining which can be used to extend the graph again, for example named entity recognition (NER) may use keywords as a context. Topic detection may also benefit from already assigned keywords, journals or author information. The graph can also be extended by knowledge discovery processes, for example finding parameters of a clinical trial, progression within electronic health records, etc. In any case new context information will be added to the initial graph and improve the input of further algorithms.

the context of knowledge graphs. We demonstrate this novel approach by applying it to common data sources. After that, we will give a detailed list of issues that have to be addressed.

I. Preliminaries

We define knowledge graphs $G = (E, R)$ where the set of nodes E consist of entities $e \in E$ coming from a formal structure like an ontology $E_i = (V(E_i), R(E_i))$. E is a union of ontologies $E = \{E_1, ..., E_n\}$. The relations $r \in R$ can be ontology relations, thus in general we can say every ontology E_i which is part of the data model is a subgraph of G which means $E_i \subseteq G$. In addition, we allow inter-ontology relations between two nodes e_1, e_2 with $e_1 \in E_1$, $e_2 \in E_2$ and $E_1 \neq E_2$. More general we define $R = \{R_1, ..., R_n\}$ as list of either inter-ontology and intra-ontology relations. Both E as well as R are finite discrete spaces.

Every entity $e \in E$ may have some additional meta information which need to be defined with respect to the application of the knowledge graph. For instance there might be several node sets (some ontologies, some document spaces (patents, research data, ...), author sets, journal sets, ...) $E_1, ..., E_n$ so that $E_i \subset E$ and $E = \cup_{i=1,...,n} E_i$. The same holds for R where several context relations might come together like "is cited by", "has annotation", "has author", "is published in", etc.

We define a finite, discrete set $C = \{c_1, ..., c_m\}$ of contexts C_i. Every node $e \in G$ and every edge $r \in R$ may have one ore more contexts $c \in C$ denoted by $con(e)$ or $con(r)$. It is also possible to set $con(e) = \emptyset$. Thus, we have a mapping $con : E \cup R \to \mathcal{P}(C)$ to the power set of C. If we use a quite general approach towards context, we may set $C = E$. Thus, every inter-ontology relation defines context of two entities, but also the relations within an ontology can be seen as context.

Every node set $E_i \in \{E_1, ..., E_n\}$ induces a subgraph $G[E_i] \subset G$. With $G^c[E_i] = G[E_i] \cup N(E_i)$ we denote the extended context subgraph which also contains the neighbours $N(E_i)$ of each node $e \in E_i$ in G, which is the context of that node. For a graph drawing perspective, if $G^c[E_i]$ defines a proper surface, we can think about a graph embedding of another subgraph $G^c[E_j]$ on $G^c[E_i]$.

We can create the metagraph $M = (C, R')$ of these contexts. Each context is identified by a node in M. If there is a connection in G between two contexts, we add an edge $(c_1, c_2) \in R'$. This means if $\exists (v_1, v_2) \in R : \; c_1 \in con(v_1), c_2 \in con(v_2) \Rightarrow (c_1, c_2) \in R'$ or $\exists (v_1, v_2) \in R : \; c_1 \in con((v_1, v_2)), c_2 \in con(v_2) \Rightarrow (c_1, c_2) \in R'$ or $\exists (v_1, v_2) \in R : \; c_1 \in con(v_1), c_2 \in con((v_1, v_2)) \Rightarrow (c_1, c_2) \in R'$. See figure 2 for an illustration.

Adding edges between the knowledge graph G or a subgraph $G' = (E', R') \subseteq G = (E, R)$ and the metagraph M in $G \cup M$ will lead to a novel graph. This can be either seen as inverse mapping $con^{-1}(G')$ or as the hypergraph $\mathcal{H}(G') = (X, \hat{E})$ given by

$$X = E' \cup G^c[E_i]$$

$$\hat{E} = \{\{e_i, e \forall e \in N(e_i)\} \forall e_i \in X\}$$

This graph can be seen as an extension of the original knowledge graph G' where contexts connect not only to the initial nodes, but also every two nodes in G' are connected by a hyperedge if they share the same context. See figure 3 for an illustration.

If $C = E$, this will lead to new edges in G enriching the original graph. This step should be done after every additional extension to the graph G. Thus we need to update both G as well as M.

We will denote this hypergraph H on a knowledge graph G and a metagraph M with $H_{G|M}$. We might also add multiple metagraphs M_1 and M_2 which will be denoted by $H_{G|M_1,M_2}$.

This graph can be seen as an enrichment of the original knowledge graph G with contexts. It can be used to answer several research questions and can be utilized to find graph-theoretic formulations of research questions.

If the mapping con is well defined for the domain set the Graph H can be generated in polynomial time. Since this is in general not the case, this usually contains a data or text mining task to generate contexts from free texts or knowledge graph entities. With respect to the notation described in [7] this problem p can be formulated as

$$p = \mathbb{D}|R|\mathbf{f} : \mathbb{D} \to \mathbb{X}|err|\emptyset \tag{1}$$

Here, the domain set $\mathbb{D}$ is explicitly given by $\mathbb{D} = G$ or – if additional full-texts $\hat{D}$ supporting the knowledge Graph G exist – $\mathbb{D} = \{G, \hat{D}\}$. In our case the domain subset $R = \mathbb{D}$. In this case we need to find a description function $f : \mathbb{D} \to \mathbb{X}$ with a description set $\mathbb{X} = C$ which holds all contexts. To find relevant contexts we need an error measure $err : \mathbb{D} \to [0, 1]$.

We have to consider several research questions. First of all: What are meta information that can be used to generate a context for a new metagraph? Good candidates are authors, citations, affiliation, journal, and MeSH-terms or rather keywords since they are available in most databases. We also need to discuss text mining results like NER, relationship mining etc. Having more general data like study data, genomics, images, etc. we might also consider side effects; disease labels, population labels (male; female; age; social class; etc.). Here we show a proof of concept for less complex text mining meta data. See figure 1, which describes the process of starting with a simple document graph that can be extended with more context data from text mining. We discuss this in more detail within the next section.

The further research questions address the application of this novel approach for both biomedical research as well as text classification and clustering, NLP and knowledge discovery, also with focus on Artificial Intelligence (AI). How can we use the context metagraph to answer biomedical scientific questions? What can we learn from connections between contexts and how do they look like in the knowledge graph? How can we use efficient graph queries utilizing the context? It may also be useful to filter paths in the knowledge graph according to a given context or to generate novel visualizations. A possible question might be to learn about mechanisms linked to co-morbidities or mechanisms being contextualized by drug information. The meta-graph may also contain information about cause-and-effect relationships in the knowledge graph that are "valid" in a biomedical sense under certain conditions. In addition, a contextualization-based on demographic information or polypharmacy information. We will discuss several use cases within the last section.

II. Method and Practical Application

The following software was written in Java using Spring Boot (see http://spring.io/projects/spring-boot) and Spring Data (see https://spring.io/projects/spring-data) and integrated in our SCAIView microservice architecture, see [8]. The database backend is a graph database running Neo4j (see https://neo4j.com/).

We will illustrate the following methods with example runs on MedLine and Pubmed data. Both sources are already included in the SCAIView NLP-pipeline. PubMed contains 29 million abstracts from biomedical literature, PMC about 4 million full-text articles.

A. Creating a document and context graph with basic context extraction

The initial step of creating a document and context graph with basic context extraction needs a basic definition of entity sets $E_1, ..., E_n$ and their relations.

The articles and abstracts from PubMed and PMC already come with a lot of contextual data. We may set $E_{Document}$ as the document set containing nodes, each representing one document. In addition, we may add a set $E_{Source} = \{\text{PubMed, PMC}\}$ as the source of a document. Thus, each document can be interpreted as context of a data source.

All meta data are stored in new node sets. E_{Author} stores the set of authors, $E_{Affiliation}$ their affiliation which is again context for the authors. Another relevant context is the publisher, in our case $E_{Journal}$. PubMed stores several classes, for example Books and Documents, Case Reports, Classical Article, Clinical Study, Clinical Trial, Journal Article, Review etc. We store this in $E_{PublicationType}$.

Another important context is $E_{Annotation}$ storing all kind of annotations like named entities or keywords, which come from the MeSH tree, see [9] and https://www.nlm.nih.gov/mesh/intro_trees.html. Thus, $E_{MeSH} \subset E_{Annotation}$ already comes with a hierarchy and edges R_{MeSH}. The value of MeSH terms and their hierarchy for knowledge extraction was shown in several recent studies like [10]. We will discuss the value of MeSH as controlled vocabulary within the next section. See figure 4 for an illustration of a single document.

All other relations can be added between the sets E_i, for example $R_{isCoAuthor}$, $R_{hasAffiliation}$, etc. With these information given it is – from an algorithmic point of view – quite easy to add all context relations like $R_{hasDocument}$, $R_{isAuthor}$, $R_{hasAnnotation}$, $R_{hasCitatio}$ etc. Edges must also store additional provenance information. See figure 5 for an illustration.

B. Extending the knowledge graph using NLP-technologies

The initial knowledge graph can be extended by NLP-technologies.

Terminologies and Ontologies are a widely considered topic in research during the last years. They play an important role in

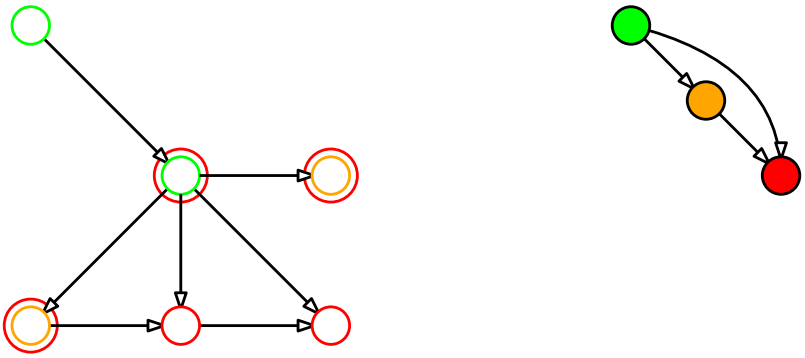

Fig. 2: Illustration of a *knowledge graph with context* (left). The context is illustrated by colors surrounding nodes. At the right the corresponding *context metagraph*. Every context in the knowledge graphs refers to a node in the metagraph and the edges describe if in the original graph a edge from one context to the next exist. Contexts may also be added to edges.

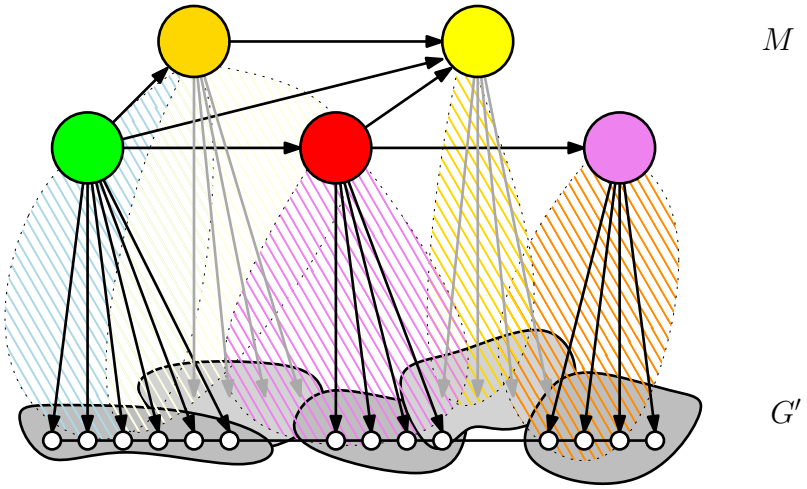

Fig. 3: This figure describes the hypergraph $\mathcal{H}(G') = (X, \hat{E})$ between the context metagraph M and the original knowledge graph G or a subgraph $G' \subset G$. This graph is sorted by contexts. The hyperedges, illustrated by sets and indicated by non-hyperedges, connect nodes with context, but also nodes with the same context.

data and text mining as well as knowledge representation in the semantic web. They become more and more important since data provider publish their data in a semantic web formats, namely RDF ([11]) and OWL ([12]), to increase integratability. The term *terminology* refers to the SKOS meta-model [13] which can be summarized as concepts, unit of thoughts which can be identified, labeled with lexical strings, assigned notations (lexical codes), documented with various types of note, linked to other concepts and organized into informal hierarchies and association networks, aggregated, grouped into labeled and/or ordered collections, and mapped to concepts. Several complex models have been proposed in literature and have been implemented in software, see [14]. *Controlled Vocabularies* contain lists of entities which may be completed to a *Synonym Ring* to control synonyms. *Ontologies* also present properties and can establish associative relationships which can also be done by *Thesauri* or *Terminologies*. See [15] and [16] for a complete list of all models.

Here we define Terminologies similar to Thesauri as a set of concepts. They form a *DAG* (Directed Acyclic Graph) with child and parent concepts. In addition, we have an associative relation which identifies similar or somehow related concepts. Each concept has one or more labels. One of them is the preferred identifier, all others are synonyms. To sum up, using ontologies or terminologies for NER, we will have a hierarchy within this ontology. But we may not only consider ontologies and terminologies, but also controlled vocabularies like MeSH. Here we have additional annotations with a different provenance, one coming as keywords with the data, one obtained from NER.

Another example is the Alzheimer's Disease Ontology (ADO, see [17]) E_{ADO} or the Neuro-Image Terminology (NIFT, see [18]) E_{NIFT} coming with their hierarchy R_{ADO}, R_{NIFT}. The process of NER will lead to another context relation $E_{hasAnnotation}$. Since not all ontologies or teminologies are described in RDF or OBO format we have to add data from

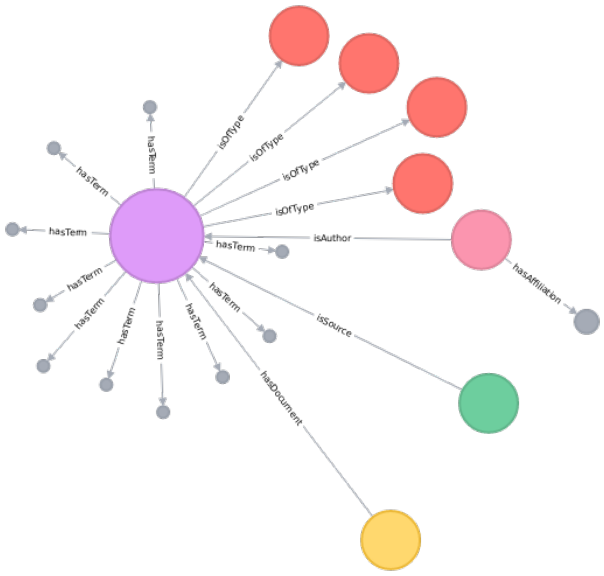

Fig. 4: This figure is an illustration of a single document within the context graph. The document node (purple) has several gray annotation nodes, four red publication type nodes, an orange author node with a gray affiliation. The source (PubMed) is annotated in a green node, the journal in a yellow node.

multiple sources. This is done by a central tool providing all ontology data.

Another context data useful for knowledge extraction are citations, thus edges $R_{hasCitation}$ between two nodes in $E_{Document}$. The data from PMC already stores citation data with unique identifiers (PubMed IDs). Some data is available with WikiData, see [19] and [20]. Other sources are rare, but exist, see [21]. Especially for PubMed a lot of research is working on this difficult topic, see for example [22].

In addition we can consider relational information between entities. For example BEL statements already form knowledge graphs of semantic triples that consist of concepts, functions and relationships [2]. To tackle such complex tasks they constantly gather and accumulate new knowledge by performing experiments, and also studying scientific literature that includes results of further experiments performed by researchers. Existing solutions are mainly based on the methods of biomedical text mining to extract key information from unstructured biomedical text (such as publications, patents, and electronic health records). Several information systems have been introduced to support curators generating these networks. BELIEF, is one workflow generated for this purpose. BELIEF build BEL like statements semi-automatically from retrieving publications from a relevant corpus generation system called SCAIView, see [23] and [24].

Figure 6 illustrates the relations *"Levomilnacipran" inhibts "BACE1"*, *"BACE1" improves "Neuroprotection"* and *"BACE1" improves "Memory"* found with relation extraction on named entities in a document. It is easy to see that context for a document is now also context for the relations and vice versa. If an entity within the relation has synonyms or is found within another document with a different context, this might lead to a deeper knowledge about the statement. Vice versa the context of the document, for example if the knowledge was found within a clinical trial, is a context to the statements.

III. Applications

We will first of all discuss some missing data or data integration problems as well as technical issues which need to be solved. Afterwards we will give an outlook on NLP-based on context information and the impact on answering semantic questions. This is highly related to the FAIRification of research data. This will lead to a short outlook on personalised medicine.

A. Missing data

We faced several issues with data integration and missing data. For example some publishers used OCR technologies to convert PDF documents in XML structures. These were usually problematic to process because some fields were missing or wrongly filled.

We have not yet worked on the problem of author and affiliation disambiguation. This is still a widely discussed topic, see [25]. An interesting novel approach – also based on Neo4j database technology – was introduced in [26]. The authors used topological and semantic structures within the graph for author disambiguation. Thus, we plan to integrate state-of-the-art technologies.

In addition performance is a major problem, and the main cause of latency for request. Thus, we had serious problems integrating this framework in our microservice architecture, see [8]. There are several possible explanations for this result, both on technical as well as implementation side. Thus, an important finding was that the storing and retrieval of large knowledge graphs did work. Not surprisingly, for giant and very dense knowledge graphs we need to find another solution. We could either improve the database backend by establishing a polyglot persistence architecture or use existing graph databases like Cray Graph Engine, see [27]. This choice has important implications for the further developing of this architecture, for example SPARQL has more limitations than Cypher. This is an important issue for future research.

However, these results were very encouraging and we will discuss some more topics for further research.

B. Context-based NLP

This novel system extends our knowledge and the availability of context data. Context data is a very important foundation for text mining [1]. For example, context-based NER was discussed by [28] and there is still ongoing research, for example on content-aware attributed entity embedding (CAAEE), see [29]. The key strength of our approach is that in every step of text mining and NLP all context data is available and new data will be added. Thus, this system can be used for

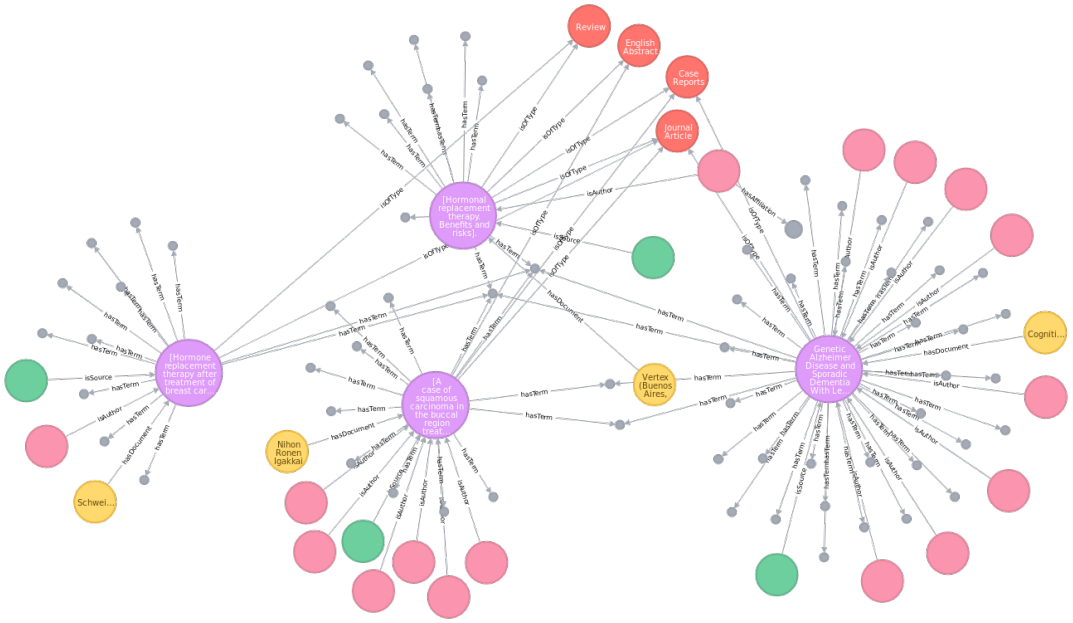

Fig. 5: This figure is an illustration of the initial document and context graph. A PubMed node is the source of document nodes (green). There are several context annotations like article type (red), keywords (gray), authors (orange) and journal (yellow). Authors have additional context (affiliations, gray).

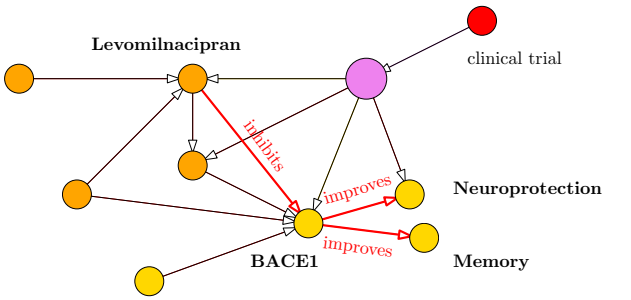

Fig. 6: This figure is an illustration of biological knowledge within the context graph. The document node (purple) has several yellow and orange annotation nodes which come from different terminologies found with NER. The relation extraction task found the relation "Levomilnacipran" inhibits "BACE1", "BACE1" improves "Neuroprotection" and "BACE1" improves "Memory". These relations are illustrated with red edges. Since the document describes a clinical trial, this is also a context for the relations as well.

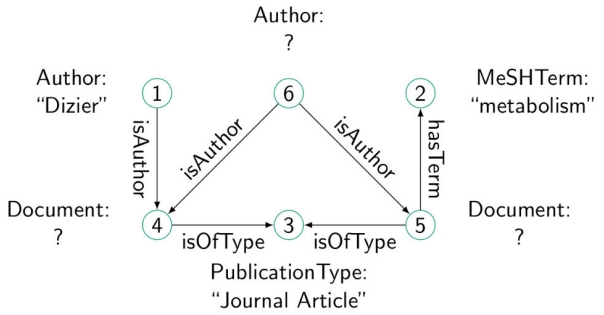

Fig. 7: This figure illustrates a more complex semantic subgraph to query the knowledge graph. We search for two documents having the same author and both of publication type "Journal Article". The first document should have an author called "Dizier", the second one a MeSH Term called "metabolism".

both building and validating Machine Learning (ML) and AI approaches.

Of course the novel context data is not only suitable for NER, but also for relation extraction. For example [30] proposed a novel approach to context-based relation extraction. Although our example is based on a small data set, the findings suggest that a lot of existing data can be utilized as context data. For example entities annotated by NER or manually curated BEL statements may be applied as context.

Thus this research has several practical applications. Firstly, it leads to validation and datasets for ML and AI approaches towards text mining. Further work needs to be done to investigate how this data can be used systematically. Secondly, it generalizes the idea of context so that it can be used for semantic questions.

C. *Answering semantic questions and FAIRification of data*

Semantic questions can be formulated as subgraph structures of the initial knowledge graphs. For example we may ask: "Which articles have been authored by Pacheco?". This will

lead to a subgraph with two nodes v_1, v_2 where $v_1 =$ Pacheco and an edge $(v_1, v_2) =$ isAuthor. We may think of much more complex examples, see figute 7 for an example.

In general these semantic subgraph queries (or: graph queries) have an input $Q = (V, E) \subset G$ and output all subgraphs $H \subset G$ with $H \simeq Q$. Thus, the problem of answering semantic questions is a generalization of the subgraph isomorphism problem. We know already subgraph isomorphism is NP-hard, see [31]. It would be interesting to find a general formulation of the generalization or restrictions that can be applied to this problem. Since Cypher already provides us with the possibility to query graph substructure, further research might explore the runtime or might lead to novel heuristics to solve this efficient.

Whilst this work did not consider the impact of novel ontologies and terminologies, it did substantiate the impact of them on context data. This is an interesting step towards the FAIRification of data. Wilkinson introduced his FAIR guiding principles in [32] referring to the findability, accessibility, interoperability, and reusability of data, especially for research data. A consequent application of the context idea leads to meta data as context on data which can afterwards be used to make meta data searchable even if the data itself is protected. Thus, the inclusion of context into an information system like SCAIView will make the data findable and accesible. In addition, if interoperable ontologies are available, this data will also be interoperable. This will already solve the three out of four issues addressed by FAIR data.

D. Perspectives for Personalised Medicine

Hypothesis generation and knowledge discovery on biomedical data are widely used in medical research and digital health. For example researchers search for genomic or molecular patterns, diagnosis or build longitudinal models. In addition, the massive data available build the basis for a multitude of predictive and personalised medicine ML and AI approaches. A reasonable approach to tackle reproducible research in predictive medicine could be to use a standardized and FAIR context graph for biomedical research data. Thus, it would be necessary to annotate not only biomedical literature but also research data like moleculare data, imaging data, genomics and electronical health records (EHR) with context information.

This information system can be used to retrieve data by context (cohort size, settings, results, ..) and by content (imaging data, genomic or moleculare measures, ...). For example, this system may answer questions like “Give me a clinical trial to reproduce my results or to apply my model” or “Give me literature for phenotype A, disease B age between C and D and a CT-scan with characteristic E”.

Here we presented a novel approach that annotates research data with context information. The result is a knowledge graph representation of data, the context graph. It contains computable statement representation (e.g. RDF or BEL). This graph allows to compare research data records from different sources as well as the selection of relevant data sets using graph-theoretical algorithms.

IV. Conclusion and Outlook

Here we discussed a proof-of-concept of a biomedical knowledge graph combining several sources of data as context to each other. We processed data from PubMed and PMC. This initial knowledge graph was extended with results from text mining and NRL-tools already included in our software. Thus, we were able to provide both small datasets as well as large collections of data.

We faced several issues with data integration and missing data, for example because the input data had a bad quality. In addition we have not yet worked on the problem of author and affiliation disambiguation. The directly leads to the question how our approach can be evaluated. For every kind of input data another evaluation method needs to be established. Without this, the quality of the knowledge graph is directly linked to the quality of input data. Before establishing a productive system, this question needs to be properly addressed.

We introduced several applications, for example context-based NLP, answering semantic questions and FAIRification of data, perspectives for Personalised Medicine. The generalisability of these ideas is subject to certain limitations. For instance, the question of interoperable ontologies or ontologies covering the issues of interoperability of data is still not examined. In addition, there is still no FAIR-data information system available.

This has thrown up many questions in need of further investigation. Nevertheless, it is not keen to make an outlook on the impact of such a FAIR and semantic information system and data structure on context data for personalised medicine.

V. Acknowledgments

We thank Tim Steinbach for providing some illustrations to this work. In addition, we thank Marc Jacobs and Alexander Esser for carefully revising the manuscript.

References

[1] C. C. Aggarwal and C. Zhai, “An introduction to text mining,” in *Mining text data*. Springer, 2012, pp. 1–10.

[2] J. Fluck, A. Klenner, S. Madan, S. Ansari, T. Bobic, J. Hoeng, M. Hofmann-Apitius, and M. Peitsch, “Bel networks derived from qualitative translations of bionlp shared task annotations,” in *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, 2013, pp. 80–88.

[3] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, “Gene ontology: tool for the unification of biology,” *Nature genetics*, vol. 25, no. 1, p. 25, 2000.

[4] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda *et al.*, “Drugbank 5.0: a major update to the drugbank database for 2018,” *Nucleic acids research*, vol. 46, no. D1, pp. D1074–D1082, 2017.

[5] K. Khan, E. Benfenati, and K. Roy, “Consensus qsar modeling of toxicity of pharmaceuticals to different aquatic organisms: Ranking and prioritization of the drugbank database compounds,” *Ecotoxicology and environmental safety*, vol. 168, pp. 287–297, 2019.

[6] C. Haupt, P. Groth, and M. Zimmermann, “Representing text mining results for structured pharmacological queries,” *ISWC*, 2011.

[7] J. Dörpinghaus, J. Darms, and M. Jacobs, “What was the question? a systematization of information retrieval and nlp problems.” in *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2018.

[8] J. Dörpinghaus, J. Klein, J. Darms, S. Madan, and M. Jacobs, "Scaiview – a semantic search engine for biomedical research utilizing a microservice architecture," in *Proceedings of the Posters and Demos Track of the 14th International Conference on Semantic Systems - SEMANTiCS2018*, 2018.

[9] F. B. Rogers, "Medical subject headings," *Bulletin of the Medical Library Association*, vol. 51, pp. 114–116, 1963.

[10] H. Yang and H. Lee, "Research trend visualization by mesh terms from pubmed," *International journal of environmental research and public health*, vol. 15, no. 6, p. 1113, 2018.

[11] R. Cyganiak, D. Wood, and M. Lanthaler, "RDF 1.1 concepts and abstract syntax," W3C, W3C Recommendation, Feb. 2014, http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.

[12] P. Patel-Schneider, S. Rudolph, M. Krötzsch, P. Hitzler, and B. Parsia, "OWL 2 web ontology language primer (second edition)," W3C, Tech. Rep., Dec. 2012, http://www.w3.org/TR/2012/REC-owl2-primer-20121211/.

[13] E. Summers and A. Isaac, "SKOS simple knowledge organization system primer," W3C, W3C Note, Aug. 2009, http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/.

[14] M. Zeng, M. Hlava, J. Qin, G. Hodge, and D. Bedford, "Knowledge organization systems (kos) standards," *Proceedings of the Association for Information Science and Technology*, vol. 44, no. 1, pp. 1–3, 2007.

[15] "Guidelines for the construction, format, and management of monolingual controlled vocabularies," National Information Standards Organization, Baltimore, Maryland, U.S.A., Standard, 2005.

[16] M. Zeng, "Knowledge organization systems (kos)," vol. 35, pp. 160–182, 01 2008.

[17] A. Malhotra, E. Younesi, M. Gündel, B. Müller, M. T. Heneka, and M. Hofmann-Apitius, "Ado: A disease ontology representing the domain knowledge specific to alzheimer's disease," *Alzheimer's & Dementia*, vol. 10, no. 2, pp. 238 – 246, 2014.

[18] A. Iyappan, E. Younesi, A. Redolfi, H. Vrooman, S. Khanna, G. B. Frisoni, and M. Hofmann-Apitius, "Neuroimaging feature terminology: A controlled terminology for the annotation of brain imaging features," *Journal of Alzheimer's Disease*, vol. 59, no. 4, pp. 1153–1169, 2017.

[19] J. Voß, "Classification of knowledge organization systems with wikidata." in *NKOS@ TPDL*, 2016, pp. 15–22.

[20] D. Vrandečić, "Toward an abstract wikipedia," in *31st International Workshop on Description Logics (DL)*, ser. CEUR Workshop Proceedings, M. Ortiz and T. Schneider, Eds., no. 2211, Aachen, 2018. [Online]. Available: http://ceur-ws.org/Vol-2211/#paper-03

[21] A. Oßwald, J. Schöpfel, and B. Jacquemin, "Continuing professional education in open access. a french-german survey," *LIBER Quarterly. The journal of the Association of European Research Libraries*, vol. 26, no. 2, pp. 43–66, 2015.

[22] A. Volanakis and K. Krawczyk, "Sciride finder: a citation-based paradigm in biomedical literature search," *Scientific reports*, vol. 8, no. 1, p. 6193, 2018.

[23] S. Madan, S. Hodapp, P. Senger, S. Ansari, J. Szostak, J. Hoeng, M. Peitsch, and J. Fluck, "The bel information extraction workflow (belief): evaluation in the biocreative v bel and iat track," *Database*, vol. 2016, 2016.

[24] S. Madan, J. Szostak, J. Dörpinghaus, J. Hoeng, and J. Fluck, "Overview of bel track: Extraction of complex relationships and their conversion to bel," *Proceedings of the BioCreative VI Workshop*, 2017.

[25] J. Kim, "Correction to: Evaluating author name disambiguation for digital libraries: a case of dblp," *Scientometrics*, vol. 118, no. 1, pp. 383–383, 2019.

[26] V. Franzoni, M. Lepri, and A. Milani, "Topological and semantic graph-based author disambiguation on dblp data in neo4j," *arXiv preprint arXiv:1901.08977*, 2019.

[27] C. D. Rickett, U.-U. Haus, J. Maltby, and K. J. Maschhoff, "Loading and querying a trillion rdf triples with cray graph engine on the cray xc," in *Cray User Group*, 2018.

[28] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[29] D. Cai and G. Wu, "Content-aware attributed entity embedding for synonymous named entity discovery," *Neurocomputing*, vol. 329, pp. 237–247, 2019.

[30] P. Prajapati and P. Sivakumar, "Context dependency relation extraction using modified evolutionary algorithm based on web mining," in *Emerging Technologies in Data Mining and Information Security*. Springer, 2019, pp. 259–267.

[31] S. A. Cook, "The complexity of theorem-proving procedures," in *Proceedings of the third annual ACM symposium on Theory of computing*. ACM, 1971, pp. 151–158.

[32] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, 2016.