Semantic Knowledge Graph Embeddings for biomedical Research: Data Integration using Linked Open Data

Jens Dörpinghaus[1,2], Marc Jacobs[1]

[1] Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, Sankt Augustin, Germany
[2] jens.doerpinghaus@scai.fraunhofer.de

Abstract. Knowledge Graphs are becoming a key instrument for biomedical knowledge discovery and modeling. These approaches rely on structured data, e.g. about related proteins or genes, and form cause-and-effect networks or – if enriched with literature data and other linked data sources – knowledge graphs. A key aspect of analysis on these graphs is the missing context. Here we present a novel semantic approach towards a context enriched Knowledge Graph for biomedical research utilizing data integration with linked data. The result is a general graph concept that can be used for graph embeddings in different contexts or layers.

1 Introduction

Biological and medical researchers considering computational approaches rely on structured data, e.g. about related proteins or genes, see [9]. Cause-and-effect networks are a special subtype of more general Knowledge Graphs. In principle, the integration of external data sources and manual curated data is key. Although several commercial solutions exist, Fakhry et al. state, that the "adoption and extension of such methods in the academic community has been hampered by the lack of freely available, efficient algorithms and an accompanying demonstration of their applicability using current public networks" [4].

This and the emerging improvements on large-scale Knowledge Graphs and machine learning approaches are the motivation for our novel approach on semantic Knowledge Graph embeddings for biomedical research utilizing data integration with linked open data. Several similar approaches (often in the context of drug-repurposing) have been described like Bio2RDF [2], hetionet [6], or Open PHACTS [5]. Our approach is more focussed on integrating the literature itself in a FAIR [10] and open knowledge graph which is also accessible from public a public resource: SCAIView[3]. SCAIView is an information retrieval system that allows semantic searches in large textual collections by ontological representations of automatic recognized biological entities [7].

[3] https://www.scaiview.com/

Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-2451/paper-10.pdf

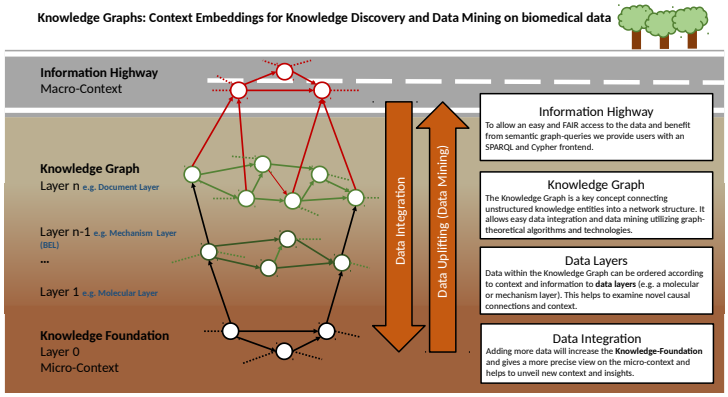

Fig. 1. Illustration of the *knowledge graph embedding* between different layers. Here, every layer corresponds to a context defining new contexts on several other layers. Thus layers and contexts are flexible and can be defined in a feasible way for every application.

The basis for generating our large-scale Knowledge Graph representation is the biomedical literature, e.g. MedLine and PubMed[4]. These articles or abstracts are the source for biological relations mentioned above. In addition, meta information like authors, journals, keywords (so called MeSH-Terms, Medical Subject Headings), etc. are freely available. Ontologies can be used to contextualize entities in the Knowledge Graph providing biological or medical relations (cf. [5]). Every ontology will form another knowledge (sub-)graph.

Using methods of natural language processing (NLP) and text mining, we can combine and link these knowledge graphs to a giant and very dense new knowledge graph. This will meet a very general definition of *context*. We can see every knowledge (sub-)graph as context to another. Biological expressions are context of the corresponding literature, authors are context of a text, named entities from ontologies found in a text are context to it or to the corresponding biological expressions.

Our overarching integration schema is based on the Biological Expression Language[6] 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. Thus they can be easily added to a knowledge graph representing another layer or context. An example for a large Alzheimer network can be found in [8].

In the next section we describe the novel concept of semantic graph embeddings within large-scale Knowledge Graphs. We will present several use-cases

[4] See https://www.ncbi.nlm.nih.gov/pubmed/.
[5] OLS, https://www.ebi.ac.uk/ols/index
[6] BEL, www.openbel.org

and application examples as well as the semantic interoperability layer using RDF and SPARQL.

2 Knowledge Graph architecture

A *Knowledge Graph* is a systematic way to connect information and data to represent common knowledge. As described above, the context is the most important topic to generate knowledge or even wisdom.

We define knowledge graphs $G = (E, R)$ with entities $e \in E$ coming from a formal structure like an ontology O, see [1] and [11]. The relations $r \in R$ can be ontology relations, thus in general we can say every ontology O which is part of the data model is a subgraph of G which means $O \subseteq G$. In addition we allow inter-ontology relations between two nodes e_1, e_2 with $e_1 \in O_1$, $e_2 \in O_2$ and $O_1 \neq O_2$. More general we define $R = \{R_1, ..., R_n\}$ as list of either ontologies, terminologies or any sort of controlled vocabulary containing relations or not.

We define contexts $C = \{c_1, ..., c_m\}$ as a finite, discrete set. Every node $v \in G$ and every edge $r \in R$ may have one ore more contexts $c \in C$ denoted by $con(v)$ or $con(r)$. It is also possible to set $con(v) = \emptyset$. Thus we have a mapping $con : E \cup R \to \mathcal{P}(C)$. 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, see figure 1 for an illustration. Here every context is identified as a layer (e.g. a document layer, a molecular layer, a mechanism layer, ...). This allows new connections between different contexts or layers: If two edges $e_1, e_2 \in R_1$ are connected and $e_1', e_2' \in R_2$ with $con(e_1) = e_1'$ and $con(e_2) = e_2'$ are not connected, we may add another edge (e_1, e_2) with provenance information that this connection comes from a different context, namely R_2. Since every layer or context can be seen as a subgraph forming a surface we can denote the relation between two layers a *knowledge graph embedding*.

It is also possible to get the context of a subgraph $R_i \subseteq G$ which can be denominated by $con(R_i)$ or with the notation of graph theory as the extended induced subgraph by the vertex set E_i from R_i given by $G^c[E_i]$. This is quite trivial if context from R_i can only be annotated to vertices in G. Then

$$G^c[E_i] = G[E_i] \cup \{(e, e') \; \forall e' \in N(e), e \in E_i\}$$

Here $con_{|E_i} = G^c[E_i]$ is the context of E_i restricted to the set of edges (relations) in the graph. The two edges e', e'' are implicitly given by this context. It is quite easy to see that the restriction on context annotated to edges makes the problem more easy from a computational perspective. Nevertheless, context on edges is needed from a real-world perspective.

The technical design was done with respect to the microservice architecture of SCAIView [3]. We offer both a REST API as well as a Java Message Service (JMS) interface. As a database backend, we used Neo4j. Here, we used Spring Data Neo4j to map objects to graphs. Thus our software can be used to perform Cypher and SPARQL queries. Data can be retrieved in JSON Graph Format or RDF format.

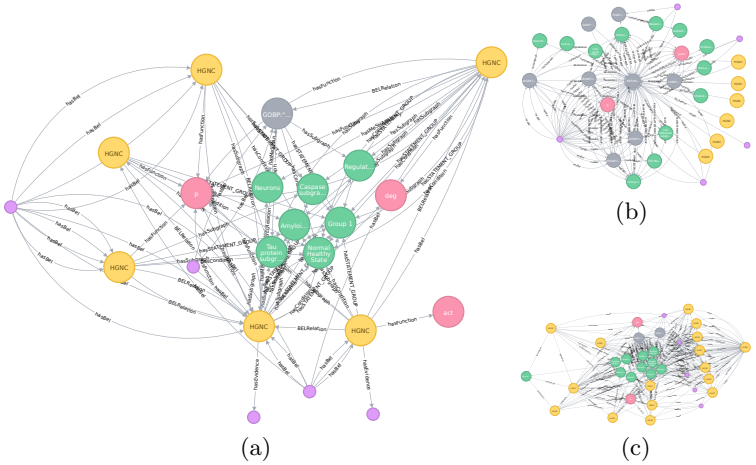

Fig. 2. (a) This is an illustration of the context found for the BEL statement `act(p(HGNC:KLC1)) => p(HGNC:MAPT)` (found on the bottom of the graph). Both HGNC terms have an evidence in two different documents (purple) and both form a relation in another document (PMID:22272245 in the middle). The green nodes form a manually curated context (e.g. "Normal Healthy State" or "Tau protein subgraph"). All HGNC entities are connected to other HGNC elements, documents and function. (b) This is an illustration of the context of a single document (purple, left). (c) This is an illustration of the context of a context (green, left).

3 Application

The initial research question was how a general context could be added to biomedical knowledge graphs to answer generic questions according to context, e.g. time, location or biological layer. We have integrated subsets of PubMed data, several ontologies like GO, HGNC, MGI and mappings, BEL networks from Parkinson's and Alzheimer's disease as well as data obtained from KEGG. See fig. 2 for some illustrations of different context layers. For example, semantic questions can be formulated as subgraph structures of the initial knowledge graphs. We may think of complex examples, e.g. "Give me all pathways from protein A to B in the context of Disease C focusing on clinical trials".

Hypothesis generation within medical research and digital health may lead to search for genomic or moleculare patterns, diagnosis or build longitudinal models which build the basis for a multitude of predictive and personalised medicine ML and AI approaches. 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.

4 Conclusion

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.

References

1. Guidelines for the construction, format, and management of monolingual controlled vocabularies. Standard, National Information Standards Organization, Baltimore, Maryland, U.S.A. (2005)
2. Belleau, F., Nolin, M.A., Tourigny, N., Rigault, P., Morissette, J.: Bio2rdf: towards a mashup to build bioinformatics knowledge systems. Journal of biomedical informatics **41**(5), 706–716 (2008)
3. Drpinghaus, J., Klein, J., Darms, J., Madan, S., Jacobs, M.: 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)
4. Fakhry, C.T., Choudhary, P., Gutteridge, A., Sidders, B., Chen, P., Ziemek, D., Zarringhalam, K.: Interpreting transcriptional changes using causal graphs: new methods and their practical utility on public networks. BMC bioinformatics **17**(1), 318 (2016)
5. Harland, L.: Open phacts: A semantic knowledge infrastructure for public and commercial drug discovery research. In: International Conference on Knowledge Engineering and Knowledge Management. pp. 1–7. Springer (2012)
6. Himmelstein, D.S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S.L., Hadley, D., Green, A., Khankhanian, P., Baranzini, S.E.: Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife **6**, e26726 (2017)
7. Hodapp, S., Madan, S., Fluck, J., Zimmermann, M.: Integration of UIMA Text Mining Components into an Event-based Asynchronous Microservice Architecture. In: Proceedings of the LREC 2016 Workshop "Cross-Platform Text Mining and Natural Language Processing Interoperability". pp. 19–23. European Language Resources Association (ELRA), Portorož, Slovenia (2016)
8. Kodamullil, A.T., Younesi, E., Naz, M., Bagewadi, S., Hofmann-Apitius, M.: Computable cause-and-effect models of healthy and alzheimer's disease states and their mechanistic differential analysis. Alzheimer's & Dementia **11**(11), 1329–1339 (2015)
9. Martin, F., Sewer, A., Talikka, M., Xiang, Y., Hoeng, J., Peitsch, M.C.: Quantification of biological network perturbations for mechanistic insight and diagnostics using two-layer causal models. BMC bioinformatics **15**(1), 238 (2014)
10. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data **3** (2016)
11. Zeng, M.: Knowledge organization systems (kos) **35**, 160–182 (01 2008)