Easing the Questioning of Semantic Biomedical Data

Arnaldo Pereira*, Rui Pedro Lopes† and José Luís Oliveira*
*DETI / IEETA, University of Aveiro, Aveiro, Portugal
†CeDRI, Polytechnic Institute of Bragança, Bragança, Portugal
Email: arnaldop@ua.pt, rlopes@ipb.pt, jlo@ua.pt

2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS) | 978-1-6654-4121-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/CBMS52027.2021.00044

***Abstract*—Researchers have been using semantic technologies as essential tools to structure knowledge. This is particularly relevant in the biomedical domain, where large dataset are continuously generated. Semantic technologies offer the ability to describe data and to map and linking distributed repositories, creating a network where the searching interface is a single entry point. However, the increasing number of semantic data repositories that are publicly available is creating new challenges related to its exploration. Despite being human and machine-readable, these technologies are much more challenging for end-users. Querying services usually require mastering formal languages and that knowledge is beyond the typical user's expertise, being a critical issue in adopting semantic web information systems. In particular, the questioning of biomedical data presents specific challenges for which there are still no mature proposals for production environments. This paper presents a solution to query biomedical semantic databases using natural language. The system is at the intersection between semantic parsing and the use of templates. It makes it possible to extract information in a friendly way for users who are not experts in semantic queries.**

***Index Terms*—Semantic Data, Semantic Web, Knowledge Graphs, Question Answering**

I. Introduction

The digitization of science in all research institutions has transformed science into a set of data-driven activities, enabling the exponential advancement of human knowledge [1]. This deluge of digital records resulted in numerous data repositories in the most varied formats, from simple spreadsheets to sophisticated databases. This situation made the reuse of data a challenge, emphasizing cases in the long tail of science where information exists closed and accessible only to the research group's elements that produced the data [2]. In the case of biomedical sciences, we find that the wide variety of repositories responds to concrete needs. Some examples are the electronic health record databases [3], data resulting from genetic studies [4], the massive collections of medical images [5], or the metadata related to biobanks' description [6]. Scientific practices established that the secondary use of data benefits various health research areas, significantly impacting the population's quality of life [7]. Therefore, researchers must have access to the best tools for sharing their data with their peers for the community's benefit.

Research in information systems tried solving the integration and interoperability of data distributed on the Internet from an early age. The Semantic Web (SW) and Linked Data (LD) principles responded to those challenges, and its use gained traction in the biomedical community [8]. Semantic technologies are at the core of many systems used, for example, in areas as diverse as translational medicine, system biology, and biopharmaceutics [9]. With the SW, the structuring of knowledge domains gained a powerful tool for formalization, the Web Ontology Language (OWL), which abstractly identifies classes, properties, and individuals [10]. This approach's success catches evident in the NCBO BioPortal, where many biomedical ontologies and terminologies are available [11].

The Resource Description Framework (RDF) is the SW's data model, establishing a basic structure, the RDF triple of a subject, a predicate, and an object. This simple way of specifying semantic units of information allows capturing biomedical data's richness in a scalable way [12]. The subject-predicate-object representation, together with ontologies, enables the annotation of knowledge and the creation of semantic repositories that can be massive. It is, therefore, necessary to have tools capable of questioning this data to obtain answers and create new knowledge. The standard strategy available out-of-the-box is the use of formal languages such as SPARQL [13]. Formal languages allow a vast range of options for forming queries, structured with their logical forms. For example, in SPARQL, if we need to retrieve variables and their bindings directly, we use the SELECT clause, and to obtain a boolean indicating a matching pattern, we ASK. Despite powerful, this and other constructs are difficult to use by non-IT people, limiting such systems' adoption.

One way to overcome the difficulties presented by systems that use formal languages is by creating interfaces that allow the use of natural language. This strategy frees users from the burden of mastering logical formalism and represents an opportunity for more users to take advantage of stored knowledge. Despite the benefits that these systems promise, the technology is not yet mature enough, and there is a need to investigate new solutions [14]. This paper presents a solution to query biomedical semantic databases using natural language, building on articulating semantic parsing and templates.

We organized the rest of the paper as follows: Section II overviews the related work in question-answering over knowledge bases. We present our solution for questioning semantic data in Section III, integrated into a semantic data creation tool. In Section IV, we use the tool to transform and explore data of patients with Huntington disease. Finally,

2372-9198/21/$31.00 ©2021 IEEE
DOI 10.1109/CBMS52027.2021.00044

Section V rounds up the paper with conclusions.

II. Question-Answering over Knowledge Bases

Generally, we call question-answering (QA) systems those interfacing databases through natural language (NL) interfaces. The goal is to obtain precise information supported by the data without using formal query languages. The implementation of these systems for the most varied data types has been investigated, considering the questioned data's specificities. Thus, some solutions specialize in conventional relational databases and other questions unstructured data such as text corpus [15]. In addition to these, a particular set of linguistic interfaces aims to take advantage of information residing in semantic databases. Sharing similar Natural Language Processing (NLP) challenges with the first types of systems, they nevertheless present particularities deserving to be highlighted [16]. When the way the entities in the question in NL are diverse from the forms used in the knowledge base (KB), we are in a lexical gap (e.g. "the King", in the NL question vs. dbp:Elvis_Presley, in DBpedia). The fact that the same phrasal name can represent several entities gives rise to ambiguity (homographs, e.g. money bank vs. river bank). Also problematic in specific contexts is the processing of complex questions that ask for aggregated, filtered, or ordered outputs. Finally, multilingualism refers to using the same interface to ask questions in several NLs and/or multilingual KB. Several proposals have emerged to tackle the enunciated difficulties grouped into the four architectural styles described in the following subsections.

A. Semantic Parsing Pipelines

The most common QA systems process data from input to output sequentially. The information passes through a pipeline's elements and transforms until it reaches a logical form digestible by the conventional SPARQL query engine. The typical architecture for this type of solution is shown in Figure 1 and consists of several blocks, commonly called filters.

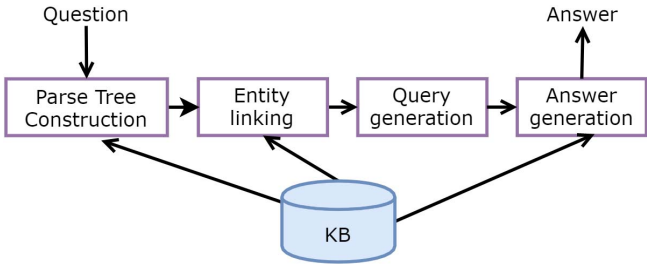

Fig. 1. Semantic parsing pipeline. The NL question is processed sequentially until a formal query is produced to obtain the answers.

Several architecture blocks correspond to known NLP elements, and many implementations are available to build tailor-made solutions. When creating the parse tree, we usually do tokenization, named-entity recognition (NER), part-of-speech (POS) tagging, and dependency parsing [17]. This way of doing, system improvements can emerge from improving particular components.

The next transformation is entity linking (EL) [18]. Although we have good solutions for constructing the parse tree for EL, demanding challenges arise when dealing with lexical gaps and ambiguities. As Ruseti et al. did, we can use an ontology to reduce ambiguity [19], but often none is available. Once the EL process is closed, a final module is responsible for transforming the parse tree with the entities and relations correctly linked into a SPARQL query.

B. Subgraphs Matching

One way to avoid difficulties with the semantic pipeline's last filters is to replace them with an architectural block for constructing subgraphs, as depicted in Figure 2.

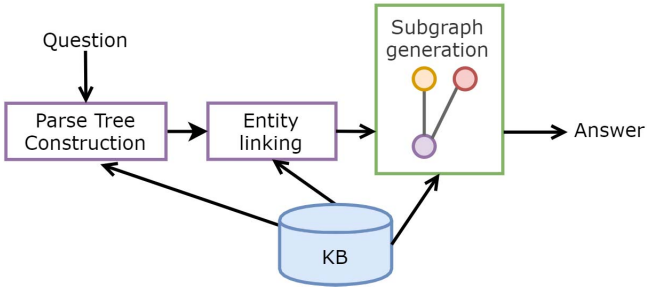

Fig. 2. Subgraphs matching. The generation of a subgraph replaces the query generation module.

Usually, this kind of solution builds upon realising that executing a formal query is equivalent to finding a subgraph [20]. Beyond this observation, it is possible to construct the answer to a question by navigating the semantic graph nodes to collect triple candidates for the final solution. Therefore, we are dealing with a search problem in a space that can be prohibitively large without considering appropriate heuristics [21]. At the end of the process, we need a strategy for selecting the most likely response.

C. Template-based QA

When looking for the answer to complex questions, the previous systems are not the most suitable. The challenges posed by the lexical gap and ambiguity cannot always be solved satisfactorily by strict semantic pipelines. The possibility of using templates allows a more accurate operation in fighting these problems [22]. A template is a query skeleton with an arbitrary degree of complexity, fitting the KB, and has slots to fill with information from entities and relations. Figure 3 outlines this type of solutions.

The creation of templates is performed offline, analysing the questions to be asked and the KB data. Solutions with a manual annotation component are common, being an obvious limitation. To have more templates is better, but the quality is also essential. Therefore, for fully automatic template generation systems, we carry on carefully. One way is to use textual information that extends the KB [23]. The online phase is easy to describe: a question is matched with a template to produce a logical form.

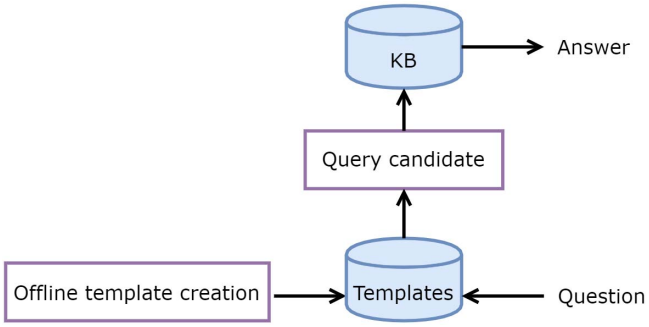

Fig. 3. Template-based QA. NL questions are mapped into pre-existing templates to be transformed into a formal query.

D. QA based on Information Extraction

When we proceed to the direct extraction of triples, we are in the presence of information extraction systems were we completely bypass the creation of a logical form. The use of machine learning techniques to create vector representations is usual (see Figure 4).

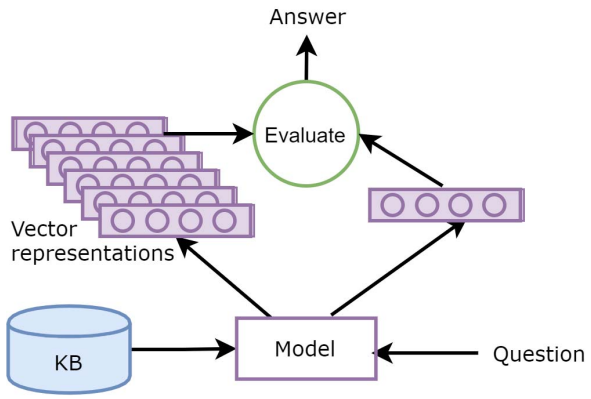

Fig. 4. QA based on information extraction. Answers are obtained directly without using a formal query.

Multiple examples in the literature rely on neural networks, as Lukovnikov et al. have done, with a character-level question encoder to handle new and rare words on the fly [24].

III. SCALEUS-FD for Question-Answering

SCALEUS-FD is a semantic web tool developed to allow data integration [25], and it is available as open-source at https://github.com/bioinformatics-ua/scaleus-fair. Quickly, we can list some of its main features:

- Very easy to deploy and start using;
- Ontology-independent;
- RDF resource loading (.ttl, .rdf, .owl, .nt, .jsonld, .rj, .n3, .trig, .trix, .trdf, .rt);
- Supports importing data from spreadsheets (.xlsx, .xls, .ods);
- Support for multiple datasets;
- Text search;
- SPARQL queries;
- Query federation to the available data;
- Inference support;
- Metadata creation allowing search engine indexing;
- Web services API.

The application offers semantic data for remote access allowing indexation by search engines crawling Data Catalog Vocabulary (DCAT)[1] descriptions. Figure 5 shows the software architecture.

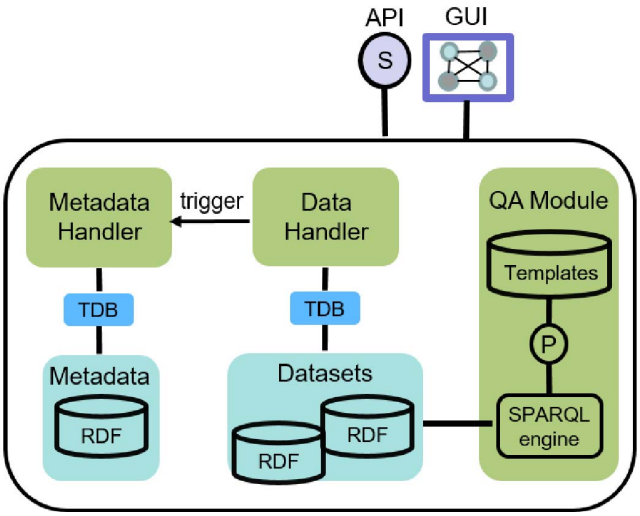

Fig. 5. SCALEUS-FD architecture.

The interface with users is via a graphical interface, and a web services API enables machine-to-machine operations. Next, in the first subsection, we outline features related to creating semantic data and metadata (Data Handler and Metadata Handler). In the second subsection, we cover the QA Module.

A. Semantic Data and Metadata Modules

The Data Handler module is responsible for transforming the information provided in a non-semantic format, such as data tables. The creation of semantic data maps the input data entities to the triples and store them in the KB. The user is free to establish a semantic scheme by creating convenient relations between data. The freedom to choose semantic prefixes is complete, and they can be created and stored for future use. Naturally, all transactions with the application's databases must ensure data integrity. The transaction database (TDB) components prevent data from being corrupted when dealing with creating, reading, updating, and deleting operations.

The metadata module ensures that data is Findable, Accessible, Interoperable, and Reusable, following the FAIR principles [26], commonly adopted in data stewardship. We ensure interoperability by using HTTP URIs to identify resources uniquely. We use the DCAT specification to characterize different layers of machine-readable metadata for describing the organizational schema catalog-dataset-distribution, which allows automatic indexation by search engines. Both data and metadata services are available through a REST API.

[1]https://www.w3.org/TR/vocab-dcat-2/

B. QA Module

The QA module allows querying the stored semantic data. On the one hand, we can operate in the traditional way by using SPARQL. This option enables advanced users to exploit all the power that a logical query language offers to construct very complex queries. On the other hand, the possibility of asking questions in natural language (in English) allows users less familiar with formal query languages to consult the knowledge stored in the KB. We integrated into the module the linguistic processing tools that allow us to do semantic parsing. Thus, the information is processed by transforming the NL question into a formal query that is then used internally to obtain the answers. But the strength of the solution is the possibility of using templates in the information retrieval process.

We can create templates in two ways. On the one hand, it is possible to provide curated lists, manually crafted. This way of doing has the advantage of capturing more precisely the users' intentions. However, it also has significant limitations. This strategy does not scale conveniently in production environments where the questioning needs give rise to new questions not covered by the previously created listings. A more efficient approach is to automate the creation of templates as carried out by the QA module (Figure 6).

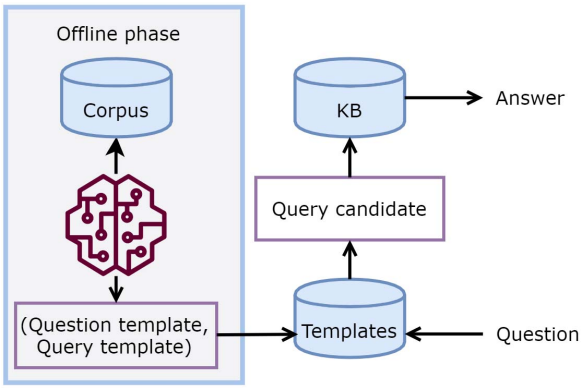

Fig. 6. Deep template-based QA.

As we can see in the figure's right branch, the system's online phase operates to transform the question in natural language into an intermediate form to pair with the appropriate template. A query is created in a formal language after filling in the slots with specific entities and relations. After this process, the final answer derives from a SPARQL query generated internally by the system.

In the offline phase, we train a deep learning model to create templates automatically. This way, we acquire more contextual information about the KB. A typical example of this procedure is the use of Wikipedia texts to expand DBpedia's knowledge. This stage is challenging since success depends on the careful choice of the set of texts we use. For instance, for a KB created by automatically extracting triples from some text corpus, this corpus can be reused to create the templates.

IV. Questioning Semantic Biomedical Data

To test the tool, we started by loading and transforming to the semantic format a spreadsheet with data from patients with Huntington disease (HD). For the sake of security and privacy, this cohort's data has been anonymized. For this example, we decided to select only a small set of headers: subject, gender, and the columns related to the Problem Behaviours Assessment (PBA-s) items [27]. We used concepts from the Dublin Core Metadata Initiative[2], FOAF Vocabulary Specification[3], and the Human Phenotype Ontology[4]. Table I shows the mapping we performed.

TABLE I
Semantic Namespace

Column	URI
subject	http://purl.org/dc/terms/identifier/
gender	http://xmlns.com/foaf/spec/#term_gender/
PBA-s Depression	https://hpo.jax.org/app/browse/term/HP:0000716/
PBA-s Irritability	https://hpo.jax.org/app/browse/term/HP:0000737/
PBA-s Psychosis	https://hpo.jax.org/app/browse/term/HP:0000709/
PBA-s Apathy	https://hpo.jax.org/app/browse/term/HP:0000741/

With the data transformed and adequately loaded, we can ask questions using a graphical interface (see Figure 7).

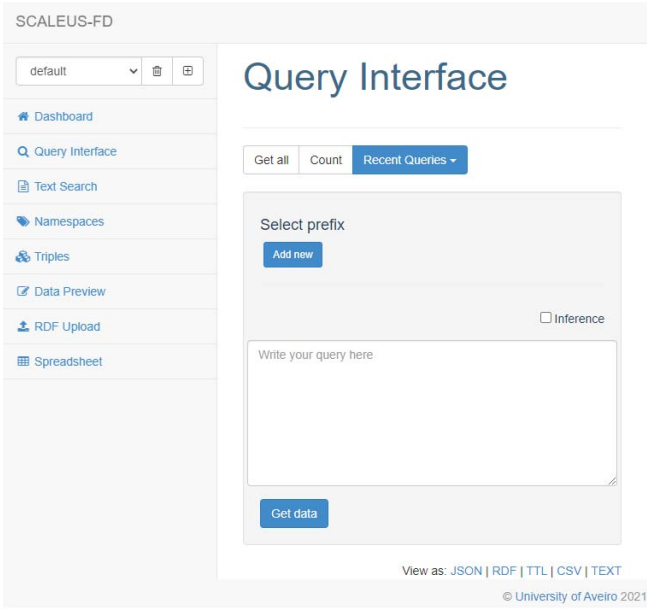

Fig. 7. QA interface.

The SPARQL queries and the NL questions use the same form for simplicity since the system recognizes the input type processing it transparently.

[2]https://dublincore.org/
[3]http://www.foaf-project.org/
[4]https://hpo.jax.org/app/

V. Conclusion

The conversion of biomedical data into a semantic format allows the sharing of relevant information between research groups. However, in addition to this essential data processing step, the systems' ability to ease retrieving information is also critical. Interfaces accepting inputs in a natural language enhance adhesion to semantic solutions. In this paper, we have proposed a tool for creating semantic data which allow us to pose questions in natural language. We believe that this tool can become part of the researchers' toolbox for their sharing of data.

Acknowledgment

FCT - Portuguese Foundation for Science and Technology supports Arnaldo Pereira (Ph.D. Grant PD/BD/142877/2018).

References

[1] E. Kolker, E. Stewart, and V. Ozdemir, "Opportunities and challenges for the life sciences community," *OMICS: A Journal of Integrative Biology*, vol. 16, no. 3, pp. 138–147, 2012.

[2] J. C. Wallis, E. Rolando, and C. L. Borgman, "If we share data, will anyone use them? data sharing and reuse in the long tail of science and technology," *PLOS ONE*, vol. 8, pp. 1–17, 07 2013.

[3] T. D. Wade, "Traits and types of health data repositories," *Health Information Science and Systems*, vol. 2, pp. 1–8, 06 2014.

[4] Y. Perez-Riverol, M. Bai, F. da Veiga Leprevost, S. Squizzato, Y. M. Park, K. Haug, A. J. Carroll, D. Spalding, J. Paschall, M. Wang, N. del Toro, T. Ternent, P. Zhang, N. Buso, N. Bandeira, E. W. Deutsch, D. S. Campbell, R. C. Beavis, R. M. Salek, U. Sarkans, R. Petryszak, M. Keays, E. Fahy, M. Sud, S. Subramaniam, A. Barbera, R. C. Jiménez, A. I. Nesvizhskii, S.-A. Sansone, C. Steinbeck, R. Lopez, J. A. Vizcaíno, P. Ping, and H. Hermjakob, "Discovering and linking public omics data sets using the omics discovery index," *Nature Biotechnology*, vol. 35, pp. 406–409, 05 2017.

[5] H. D. Tagare, C. C. Jaffe, and J. Duncan, "Medical Image Databases: A Content-based Retrieval Approach," *Journal of the American Medical Informatics Association*, vol. 4, pp. 184–198, 05 1997.

[6] G. Jacobs, A. Wolf, M. Krawczak, and W. Lieb, "Biobanks in the era of digital medicine," *Clinical Pharmacology & Therapeutics*, vol. 103, no. 5, pp. 761–762, 2018.

[7] A. G. Villanueva, R. Cook-Deegan, B. A. Koenig, P. A. Deverka, E. Versalovic, A. L. McGuire, and M. A. Majumder, "Characterizing the biomedical data-sharing landscape," *Journal of Law, Medicine and Ethics*, vol. 47, no. 1, p. 21–30, 2019.

[8] P. Sernadela, L. González-Castro, C. Carta, E. van der Horst, P. Lopes, R. Kaliyaperumal, M. Thompson, R. Thompson, N. Queralt-Rosinach, E. Lopez, L. Wood, A. Robertson, C. Lamanna, M. Gilling, M. Orth, R. Merino-Martinez, M. Posada, D. Taruscio, H. Lochmüller, P. Robinson, M. Roos, and J. L. Oliveira, "Linked registries: Connecting rare diseases patient registries through a semantic web layer," *BioMed Research International*, vol. 2017, p. 1–13, 2017.

[9] H. Chen, T. Yu, and J. Y. Chen, "Semantic Web meets Integrative Biology: a survey," *Briefings in Bioinformatics*, vol. 14, pp. 109–125, 04 2012.

[10] P. Hitzler, M. Krötzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph, "OWL 2 Web Ontology Language Primer (Second Edition)," Dec. 2012. Accessed on: Mar. 8, 2021. [Online]. Available: https://www.w3.org/TR/owl2-primer/.

[11] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen, "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications," *Nucleic Acids Research*, vol. 39, pp. W541–W545, 06 2011.

[12] B. Golshan, A. Halevy, G. Mihaila, and W.-C. Tan, "Data integration: After the teenage years," in *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, PODS '17, (New York, NY, USA), p. 101–106, Association for Computing Machinery, 2017.

[13] W3C SPARQL Working Group, "SPARQL 1.1 Overview," Mar. 2013. Accessed on: Mar. 8, 2021. [Online]. Available: https://www.w3.org/TR/sparql11-overview/.

[14] K. Affolter, K. Stockinger, and A. Bernstein, "A comparative survey of recent natural language interfaces for databases," *The VLDB Journal*, vol. 28, pp. 793–819, 08 2019.

[15] E. Dimitrakis, K. Sgontzos, and Y. Tzitzikas, "A survey on question answering systems over linked data and documents," *Journal of Intelligent Information Systems*, vol. 55, p. 233–259, 01 2020.

[16] K. Höffner, S. Walter, E. Marx, R. Usbeck, J. Lehmann, and A.-C. Ngonga Ngomo, "Survey on challenges of question answering in the semantic web," *Semantic Web*, vol. 8, pp. 895–920, 11 2017.

[17] J. Berant, A. Chou, R. Frostig, and P. Liang, "Semantic parsing on Freebase from question-answer pairs," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, (Seattle, Washington, USA), pp. 1533–1544, Association for Computational Linguistics, Oct. 2013.

[18] W. Shen, J. Wang, and J. Han, "Entity linking with a knowledge base: Issues, techniques, and solutions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 2, pp. 443–460, 2015.

[19] S. Ruseti, A. Mirea, T. Rebedea, and S. Trausan-Matu, "Qanswer - enhanced entity matching for question answering over linked data," in *CLEF2015 Working Notes*, vol. 1391, pp. 1–12, 01 2015.

[20] S. Hu, L. Zou, J. X. Yu, H. Wang, and D. Zhao, "Answering natural language questions by subgraph matching over knowledge graphs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 5, pp. 824–837, 2018.

[21] W.-t. Yih, M.-W. Chang, X. He, and J. Gao, "Semantic parsing via staged query graph generation: Question answering with knowledge base," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (Beijing, China), pp. 1321–1331, Association for Computational Linguistics, July 2015.

[22] J. Berant and P. Liang, "Semantic parsing via paraphrasing," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Baltimore, Maryland), pp. 1415–1425, Association for Computational Linguistics, June 2014.

[23] A. Abujabal, M. Yahya, M. Riedewald, and G. Weikum, "Automated template generation for question answering over knowledge graphs," in *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, p. 1191–1200, 2017.

[24] D. Lukovnikov, A. Fischer, J. Lehmann, and S. Auer, "Neural network-based question answering over knowledge graphs on word and character level," in *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, p. 1211–1220, 2017.

[25] A. Pereira, R. P. Lopes, and J. L. Oliveira, "Scaleus-fd: A fair data tool for biomedical applications," *BioMed Research International*, vol. 2020, pp. 1–8, 2020.

[26] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L. O. B. da Silva Santos, and M. D. Wilkinson, "Cloudy, increasingly fair; revisiting the fair data guiding principles for the european open science cloud," *Information Services & Use*, vol. 37, no. 1, pp. 49–56, 2017.

[27] G. McNally, H. Rickards, M. Horton, and D. Craufurd, "Exploring the validity of the short version of the problem behaviours assessment (pba-s) for huntington's disease: A rasch analysis," *Journal of Huntington's Disease*, vol. 4, no. 4, pp. 347–369, 2015.