Journal of Biomedical Informatics 137 (2023) 104272

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Querying semantic catalogues of biomedical databases

Arnaldo Pereira [a,*], João Rafael Almeida [a,b,1], Rui Pedro Lopes [c], José Luís Oliveira [a]

[a] DETI/IEETA, LASI, University of Aveiro, Aveiro, Portugal
[b] Department of Computation, University of A Coruña, A Coruña, Spain
[c] CeDRI, Polytechnic Institute of Bragança, Bragança, Portugal

ARTICLE INFO

Keywords:
Biomedical data
Knowledge bases
Semantic data
Linked data
Information extraction
Natural language interfaces
Question answering

ABSTRACT

Background: Secondary use of health data is a valuable source of knowledge that boosts observational studies, leading to important discoveries in the medical and biomedical sciences. The fundamental guiding principle for performing a successful observational study is the research question and the approach in advance of executing a study. However, in multi-centre studies, finding suitable datasets to support the study is challenging, time-consuming, and sometimes impossible without a deep understanding of each dataset.

Methods: We propose a strategy for retrieving biomedical datasets of interest that were semantically annotated, using an interface built by applying a methodology for transforming natural language questions into formal language queries. The advantages of creating biomedical semantic data are enhanced by using natural language interfaces to issue complex queries without manipulating a logical query language.

Results: Our methodology was validated using Alzheimer's disease datasets published in a European platform for sharing and reusing biomedical data. We converted data to semantic information format using biomedical ontologies in everyday use in the biomedical community and published it as a FAIR endpoint. We have considered natural language questions of three types: single-concept questions, questions with exclusion criteria, and multi-concept questions. Finally, we analysed the performance of the question-answering module we used and its limitations. The source code is publicly available at https:// bioinformatics-ua.github.io/BioKBQA/.

Conclusion: We propose a strategy for using information extracted from biomedical data and transformed into a semantic format using open biomedical ontologies. Our method uses natural language to formulate questions to be answered by this semantic data without the direct use of formal query languages.

1. Introduction

The digitisation of medical information resulted in amounts of digital health data used to support health professionals. However, this data can also be used as a powerful source of information to create new knowledge. Secondary use of data is a successful strategy for reducing costs and overcoming difficulties arising when primary data creation procedures are expensive or when target populations are small, as is the case, for example, with rare disease patients [1]. Over time, researchers worldwide have created repositories of biomedical data in various formats, such as specialised databases or simple tabular data [2]. However, the existence of this data is naturally less effective when it is not possible to share or integrate it with other data. Sharing data translates into numerous advantages for researchers. It improves data availability and linkage to other relevant sources of information. This promotes new fields of study and significantly increases the impact and recognition of research outputs [3].

We can point out different strategies to solve data sharing and interoperability problems [4]. One approach is to map the original data to a relational common data model, as advocated by international consortia such as the Observational Health Data Sciences and Informatics (OHDSI) initiative [5]. This approach focuses on agreement among domain experts on relevant concepts after systematically analysing observational data dispersed across multiple databases. In addition, a set of tools and strategies allows extracting and transforming the original data into the new format to be loaded into a database or made available as tabular data. Naturally, there is the downside that information from databases with sensitive information is somehow made available to the community, requiring extra effort to protect clinical data in data harmonisation and migration operations due to legal and ethical

* Corresponding author.
E-mail addresses: arnaldop@ua.pt (A. Pereira), joao.rafael.almeida@ua.pt (J.R. Almeida), rlopes@ipb.pt (R.P. Lopes), jlo@ua.pt (J.L. Oliveira).
[1] Equal contribution with first author to this research.

https://doi.org/10.1016/j.jbi.2022.104272
Received 10 March 2022; Received in revised form 3 November 2022; Accepted 12 December 2022
Available online 20 December 2022

constraints [6]. A strategy used for publishing the existence of these databases is based on characterising each dataset, using data aggregation and *meta*-data, and instead of releasing the databases, these characterisations are publicly available in a database catalogue. Researchers can analyse the *meta*-data and find the databases that should fit the study's needs.

There is a time end-users need to find the datasets of interest to support their study. Correctly selecting the study design and databases is essential to ensure the study's feasibility. This selection is made according to the available data that can answer research questions about the study target, considering current advances in the area. Therefore, user interface functionalities are central elements for the successful use of the system. Although logical query languages allow extracting any desired information, their handling is complex and reserved for computer specialists [7]. Contrarily, the objective is simple data access without losing power in the question formulation. A first approach to solving this problem can be the use of query builders. Using predefined options, a query builder guides users by providing question skeletons. However, this solution has the critical limitation of being closely linked to the data schema, which implies that users should at least know some details of the logical structuring of the data [8].

Question Answering over Knowledge Bases (KBQA) systems make it possible to ask questions in natural language (NL) to obtain concise answers from semantic databases, freeing users from knowledge of the data schema and formal languages [9]. These systems rely heavily on advanced Natural Language Processing (NLP) techniques and are constantly evolving to accommodate increasingly complex natural language queries. However, their use for biomedical semantic data remains challenging because of lexical ambiguity, question abstraction issues, and query generation problems [10].

In this paper, we proposed a methodology to help medical researchers to find the datasets of interest to support their studies by using an interface that receives queries in natural language. Our proposal was validated using a catalogue of Alzheimer's disease datasets and the source code is publicly available at https://bioinformatics-ua.github.io/BioKBQA/. In summary, our main contributions in this paper are the following:

- A method for converting natural language questions into a logical representation, allowing users to use natural language interfaces.
- A strategy for creating biomedical datasets semantic descriptions and promoting data findability, accessibility, interoperability, and reusability.

2. Related work

The use of semantic technologies allows researchers to share their data in a distributed and interoperable way. In this context, it is essential how we can query these data and the maturity of the available user interfaces. In addition, several life sciences communities' search for biomedical semantic datasets made it essential to create metadata catalogues related to datasets of interest.

2.1. Discovery of biomedical databases

Searching for datasets raises different challenges from those faced with current web searches [11]. When looking for datasets, users are also interested in using and retrieving data characterisations, such as the data origin, the data production date, publication formats, access policies, and the number of records, among others [12]. Other difficulties are raised by the proliferation of publishers with publishing practices outside known platforms, which does not favour finding the datasets, even if they are somewhere on the web [13]. Achieving this type of search in a similar way to that of current web search engines is still very dependent on the metadata offered by the entity that provides the dataset with crawlers only recognising some vocabularies such as Schema.org[2] [14].

Data discovery solutions must provide intuitive interfaces that allow users different ways of carrying out their searches. It is also desirable that the solutions adhere to the FAIR (Findable, Accessible, Interoperable and Reusable) guidelines. The FAIR data principles intend to ensure that humans and machines can discover and reuse data resources [15]. The key idea behind formulating these principles is to be as comprehensive as possible in summarising data custodians' best practices without committing to any implementation decisions [16]. We must assign data and metadata persistent identifiers and guarantee registration in a searchable resource. We must use relevant attributes that adhere to community standards pertinent to the domain. Data and metadata must have a formal representation using FAIR-compliant vocabularies. We must be able to retrieve data and metadata using a standardised communication protocol allowing authentication and authorisation when necessary. Finally, metadata must remain accessible even when the annotated data is no longer available.

The application of semantic technologies is at the base of several platforms. BioSharing covers life science topics related to standards, databases, and policies [17]. Also, YummyData [18] is based on Linked Data to promote the discovery of biomedical databases and the Open PHACTS Discovery Platform [19] regarding pharmacological databases. DataMed uses the DATS unified data model to allow metadata submission about datasets and provides a search engine that allows users to enter queries [20]. The EHR4CR platform integrates clinical data from several hospitals and pharmaceutical companies in seven European countries [21]. The EMIF-Catalogue is used for sharing and reusing biomedical data. Through this system, data custodians can publish and share different levels of information, while the researchers can search for databases that fulfil research requirements [22].

2.2. Managing biomedical data with semantic web technologies

Semantically organised data present a logical structure that facilitates inferring new knowledge, and we can use it directly to answer questions [23]. Therefore, it is convenient that we can store the knowledge extracted from structured or unstructured data in a Knowledge Base (KB). We can consider that the data of a KB is organised as an edge-labelled multidigraph [24]. Nodes usually represent real-world entities or quantities, and labelled arcs represent relationships between entities. Semantic web standards go further in formalising and restricting the nature of KB elements. RDF (Resource Description Framework) data consists of triples (s, p, o), where s is the subject (the resource being described), p is the predicate (the property), and o is the object (the property value) [25]. Based on this simple data model, we can build more complex models by semantic extension. As an illustration, we can point out some biomedical knowledge bases. The DrugBank[3] database combines chemical, pharmacological, and pharmaceutical information on drugs with their targets' sequence, structure, and pathways information [26]. The Universal Protein Resource (UniProt)[4] offers manually curated protein sequences and functional annotation [27].

Standard vocabularies and ontologies allow us to model shared conceptualisations of knowledge domains by establishing classes, properties, individuals, and data values [28]. We can point out some notable contributions regarding life sciences. The Human Phenotype Ontology (HPO) vocabulary describes human diseases' phenotypic abnormalities [29]. The Orphanet Rare Disease Ontology (ORDO) is a resource for annotating rare disease data that provides relationships between relevant traits, namely diseases and genes [30]. Gene Ontology (GO) describes genes considering molecular functions, cellular

[2] https://schema.org/.
[3] https://go.drugbank.com/.
[4] https://www.uniprot.org/.

components, and biological processes [31]. The ELIXIR[5] (European Life Sciences Infrastructure for Biological Information) initiative also offers an ontology repository platform [32]. Many more biomedical ontologies and terminologies are available on the BioPortal repository [33], sponsored by the National Center for Biomedical Ontology (NCBO).

Several organisations and projects dealing with biomedical data benefit from using semantic approaches. ELIXIR organisation's primary goal is to bring together life science resources across Europe. ELIXIR's activities touch five areas: 1) register and benchmark of software tools, 2) data access, 3) data interoperability, 4) cloud computing platforms, and 5) the establishment of a training community for researchers across Europe. The RD-Connect[6] initiative created an infrastructure for rare disease research to improve the analysis and sharing of genomic data, patient registries, and virtual biobanks [34]. The Biodiversity Community Integrated Knowledge Library (BiCIKL)[7] project aims to promote open science by providing access to data, tools, and services related to biodiversity research, pointing out various data linking strategies, namely using semantic technologies [35].

2.3. Natural language interfaces to semantic data

A solution for retrieving facts from a semantic database is to use semantic search engines based on keywords [36]. SANT [37] allows the publication, browsing, and querying of arbitrary RDF data. SANTé keyword-based search engine relies on building a network of terms using the values of the rdfs:label[8] property, following the formalisation of Marx *et al.* [38]. Azad *et al.* [39] proposed a system allowing users to enter the search term and to choose whether to perform a forward or a backward search. In forwarding search, the term inserted is a triple's subject, aiming to obtain triple's objects, while in backward search, we start from the object to the subjects. Another approach is Semankey [40] which creates SPARQL queries from a list of user-entered keywords. The tool pipeline consists of an entity recognition module, an ontology-based tree generator, and a query generator that uses a set of rules to translate previously produced query trees into SELECT queries with filters. However, natural language interfaces must go beyond keyword-based search, allowing the processing of more complex inputs by capturing the dependency tree of the questions or other sophisticated patterns between different lexical items [41].

We can divide the solutions for creating natural language interfaces for knowledge bases into two main groups: 1) semantic parsing and 2) information extraction. In the first group, we perform semantic parsing by applying NLP techniques intending to transform the NL question into a formal query that we use to obtain the answers, ending the process. In the second group, we find solutions to get the answers by extracting information directly from the knowledge base without creating a formal query.

Hamon *et al.* [10] answer questions about linked biomedical data by applying a multi-step method. First, they automatically annotate the NL question to identify named entities. Then we have a phase in which surface linguistic elements are linked to biomedical entities. Finally, a fixed set of rules allows us to build the desired SPARQL query. Similarly, the QuerioDALI [42] system first executes a NER and then an entity linking (EL) filter. After obtaining a set of possible results, fusion and classification are used for the final choice of answers. Bio-SODA [43] performs preprocessing for building indexes and graph schema generation on the first run of the system. A set of sequential operations is performed, mapping the question tokens to the database items, sorting the candidate tokens, building candidate query graphs, and creating the SPARQL query.

Using ontologies reduces ambiguity in identifying entities in a knowledge base. Recognising this, Ruseti *et al.* [44] map phrasal constructions to entities of an ontology using the DBpedia and Wikipedia. Yin *et al.* [45] resolve the lexical gap by considering paraphrases of the inserted question. We can use lexicons and grammars to tackle ambiguity and the lexical gap. Hakimov *et al.* [46] use a combinatorial categorical grammar that contains manually constructed lexical entities and lambda calculus expressions to build the semantic representations of possible questions. The performance of systems based on lexicons improves when they increase, as indicated by Yih *et al.* [47].

TR Discover [48] solution uses a grammar that maps first-order logical expressions to SPARQL. Dubey *et al.* [49] also propose a grammar but consider an additional normalisation step to create intermediate canonical syntactic elements representing NL questions.

In semantic parsing pipeline systems, the final phase corresponds to query generation. Although the previous process of extracting entities and linking them has been successful, this phase is not trivial. Zafar *et al.* [50] use the entities and relationships identified to generate walks through the semantic graph, considering the adjacent links within a one-hop distance. Another restriction is to consider only the tours containing the identified entities. Before creating SPARQL queries, a selection process also looks at the input question type. Abdelkawi *et al.* [51] complement this proposal by adding restrictions to allow answering ordinal and filter questions.

Modular systems are conducive to component reuse. Frankenstein [52] is a platform that makes it possible to use different alternative blocks for a given task. Modern question-answering systems allow flexible integration of specialised components. Singh *et al.* [53] consider that the choice of modules is an optimisation problem where each element is selected independently, and then the whole system's performance is evaluated. Predicting the components with better results for a given semantic data set is a supervised learning problem for these authors.

We can use modern deep learning techniques to solve many problems in KBQA systems. Dong *et al.* [54] use a multicolumn convolutional neural network that considers the answer path, answer context, and answer type. The distributed representations of the three dimensions are learned. It is possible to enrich the system by adding new aspects, paying a computational price that needs to be considered. Xu *et al.* [55] use a neural network to extract possible answers from Freebase and validate them using Wikipedia. The method also applies a set of syntactic patterns to divide more complex questions into sub-questions.

To answer simple questions, Lukovnikov *et al.* [56] rank subject-predicate pairs with a neural network that contains a nested word and encodes the questions at the character level. This approach allows dealing with new and rare surface elements without compromising semantic exploration at the word level. No semantic parsing pipeline is built, which prevents error propagation. After retraining, we can reuse the solution for new domains. The use of deep learning is affected when the amount of training data is limited, leading to overfitting [57]. To better understand the NL question, several authors such as Lukovnikov *et al.* [58], Luo *et al.* [59], and Panchbhai *et al.* [60] have been using Bidirectional Encoder Representations from Transformers (BERT). Despite continuous advances in KBQA, applying these systems to biomedical metadata remains challenging, which makes our proposal relevant.

3. Materials

Our proposal aims to add new functionality to search in natural language for semantic data. This work seeks to integrate it into a solution for creating biomedical data catalogues and an ontology repository developed by our research group in previous work.

[5] https://elixir-europe.org/.
[6] https://rd-connect.eu/.
[7] https://bicikl-project.eu/.
[8] https://www.w3.org/TR/rdf-schema/#ch_label.

3.1. MONTRA framework

In multicentre studies, there is a need to identify the best datasets to conduct a research study. With the explosion of data creation in the medical community, ideas like using catalogues to collect dataset characteristics gained momentum. Community catalogues fit into this philosophy, enabling research groups with the same interests to share metadata about their databases.

The EMIF initiative focused on creating a European Medical Information Framework to provide better healthcare using the vast amounts of biomedical data available. A web solution was thus designed to offer the EMIF Catalogue, a FAIR platform where data custodians can publish metadata about their biomedical databases with different levels of granularity [61]. This catalogue used the MONTRA framework to allow the publishing and discovery of data [22].

The MONTRA framework can create database catalogues using a data skeleton to capture the entities of interest. This skeleton can be defined by the data owners using a simple spreadsheet which is then loaded to determine the catalogue fields. The solution's architecture is flexible and allows for the integration of external components. Plugin integration can increase the basic functionality. For example, we can add a new metadata search module, improving the base search capabilities. The solution also incorporates a REST API that allows interactions with third-party software applications [62].

Search functionalities are a central aspect of a catalogue's good operation. The MONTRA platform allows users to search for datasets using forms like a query builder. We can build our survey in its simplest version by filling in a predefined set of fields. The operator AND then operates these fields. This more simplified search model only allows the construction of simple queries, which does not always serve users' interests. We also have a form with all the possible fields, with which we can build complex questions using the AND and OR operators. However, this functionality is problematic for most users as it implies thorough knowledge of the solution's metadata layer.

Using questions in natural language is an asset for users because it allows the construction of complex queries without prior knowledge of the data structure. In our proposal, MONTRA provides the base infrastructure to expose biomedical databases, that was enriched by adopting a natural language interface. The integration of our system is better described in Section 4.2, where it adopted the MONTRA principles regarding plugin integration. This is an addition to the classic form-based search methods that this framework already has.

3.2. SCALEUS-FD

A catalogue of biomedical datasets, such as those that can be built using MONTRA, provides users with a centralised access point to descriptions that help them make decisions with a profound impact on their research. Conveniently, these descriptions can be found using suitable user interfaces to facilitate this work. Mapping data in a semantic format using an ontology allows linking and relating the metadata, simplifying searching.

The management of multiple semantic datasets can be operationalised using a tool such as SCALEUS-FD. This solution allows the conversion of tabular data into semantic data. In addition to this primary function, the solution is a robust solution when used as an ontology repository. Software agents can load and access ontologies since SCALEUS-FD offers a RESTful API to perform these operations [63].

The publication of ontologies must ensure that they can be registered or indexed by search engines. Their findability is crucial for researchers to benefit from their information. In addition, we need to ensure they can be accessed using open communication protocols that allow machine-machine interactions. Data interoperability is assured when using semantic standards. As for the reuse of data, access policies must be perfectly defined and available to users. All these characteristics guarantee that the data is FAIR, as prescribed by good practices. SCALEUS-FD is an ontology repository that ensures all these desirable FAIR characteristics, as assessed by Pereira *et al.* [63] using the maturity metrics proposed by Wilkinson *et al.* [64].

When we use metadata to describe catalogues, we establish how the data can be accessed and reused. To create access points to catalogues described by metadata and allow their interoperability, they must follow a standard vocabulary such as Data Catalog Vocabulary (DCAT).[9] SCALEUS-FD uses RDFa to enable web crawlers to index DCAT annotations automatically.

Due to the high number of characteristics of each dataset fingerprint, we acknowledge that creating better ways of data search would optimise the cohort selection process. A common way researchers define cohorts is by constructing questions. Inspired by this philosophy, we created a question answering (QA) system to identify databases in the catalogue, formulating questions in natural language.

4. Methods

We propose a semantic data questioning system using natural language and its integration in a biomedical database catalogue solution (Fig. 1). The solution includes several phases, starting with the creation of lexicons of entities and relationships. These lexicons, plus a set of question-answers pairs, are used in the subsequent phases. The template generation allows for capturing the main components of natural language questions and formal language queries, while the generalisation phase makes it possible to construct a more generic base to cover other use cases. The integration of these templates in a database catalogue platform and its operation are the final steps of our pipeline and are further detailed in section 4.2.

4.1. Natural language queries over knowledge bases

Considering the pairwise disjoint sets I of IRIs, B of blank nodes, and L of literals, an RDF-Schema KB is an edge labelled multidigraph $K = (V, E^*)$ that is defined by a node set $V = V_1 \cup V_2$ with $V_1 = \mathrm{I} \cup \mathrm{B}$, $V_2 = \mathrm{I} \cup \mathrm{B} \cup \mathrm{L}$, and a labelled arc set $E^* = \{(v_1, l, v_2): v_1 \in V_1, v_2 \in V_2, l \in Lbl\}$, l being an element of the label set $Lbl = \mathrm{I} \cup \mathrm{B} \cup \mathrm{L}$. A labelled arc will commonly be called a predicate. As for their quality, nodes can be of different natures. More specifically, the set of nodes can be broken down into $C \cup In \cup \mathrm{L}$, where C is a set of classes, In is a set of class instances, and L is a set of literals. Connecting two nodes, we can have a multitude of predicates. Each pair of nodes plus the connecting predicate is called a fact. A path is a sequence $(v_0, a_1, v_1, \ldots, a_n, v_n)$, *n greater than 0*, alternating nodes (v_i, $i = 0, \ldots, n$) and labelled arcs (a_j, $j = 1, \ldots, n$). The length of a path is equal to its number of arcs. The shortest paths between two nodes are those that contain the fewest number of arcs. The smallest subgraph containing a subset N of nodes comprises all shortest paths between all pairs of nodes of N. To accommodate more complex cases, we also consider nodes representing n-ary relations, which can be coded by creating an individual that represents the relation instance itself or using an RDF vocabulary for lists.

4.1.1. Creation of lexicons

We start building two lexicons using distant supervision to use later to eliminate the ambiguity of phrasal nouns and phrasal verbs identified in the NL question. More precisely, we create a lexicon Lex_e mapping text fragments to entities and a lexicon Lex_r mapping text fragments to relations. The starting point is to annotate entities of interest on a text corpus with DBpedia Spotlight [65]. To build Lex_e, we take each (e_1, r, e_2) triple and detect, for instance, the $\langle e_1\ r$ *syntactic unit*$_1\rangle$ and $\langle$*syntactic unit*$_2\ r\ e_2\rangle$ patterns in the annotated texts, adding to the lexicon the mappings {*syntactic unit*$_1 \rightarrow e_2$, *syntactic unit*$_2 \rightarrow e_1$}. In constructing the predicate lexicon, we follow a similar principle. For this set, considering

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

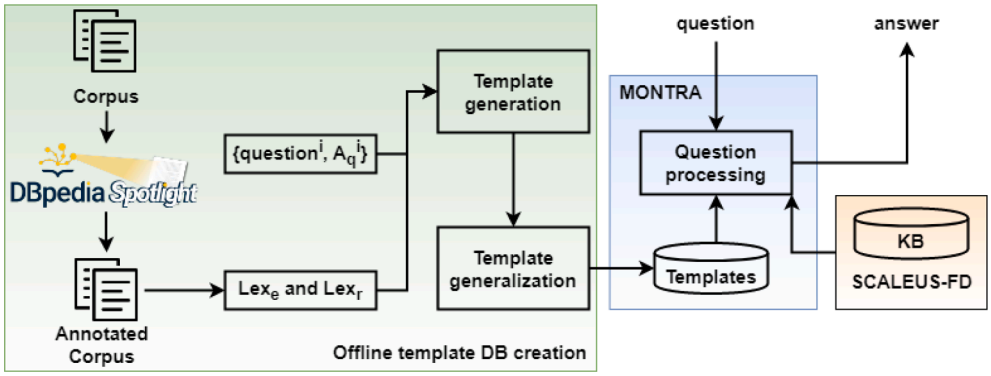

Fig. 1. General overview of the approach.

each (e_1, r, e_2) triple, we identify the patterns $\langle e_1$ *syntactic unit* $e_2\rangle$ and add to Lex_r the mapping {*syntactic unit* $\rightarrow r$}. Note that more patterns can be added later to increase the system's sensitivity.

4.1.2. Template generation

A query q is a set of triples patterns, and the answers to that query will be denoted by A_q. Templates are generated at training time to allow us to answer questions at testing (answering) time. Training stage inputs are pairs of NL questions u and its answer set A_u, being the training set $T = \{(u, A_u)\}$. We start by annotating a training NL question u with the named entities it contains and disambiguating these to KB entities using a named-entity disambiguation system, as we can see in Fig. 2.

Next, as we can see in Fig. 3, for each answer $a \in A_u$, we find the smallest subgraph of the KB that contains the entities found in the question as well as a. To this end, starting with an entity found in the question, we explore all paths of length two when the middle node is an n-ary node and paths of length one otherwise, to restrict the search space. Like Yih *et al.* [66], we assume that this subgraph captures the meaning of the question and connects it to one of its answers a. There may be multiple such graphs. Then we transform each subgraph into a backbone query $\widehat{q}$ by replacing a with the variable ?x. Note that this procedure is performed for each $a \in A_u$ for a given u, resulting in multiple queries.

Capturing the answer types given in the question is important for precision. Identifying the expected answer type of an utterance boosts the performance of QA systems. We automatically create templates that capture which phrases in the question evoke types in the query, and use the full KB type system as potential mapping targets. Starting with $\widehat{q}$ generated thus far, we connect to the answer variable node in $\widehat{q}$ one type constraint for each $c \in C$ such that the variable originates from the answer entity $a \in A_u$ and (a, *type*, c) $\in KB$ (Fig. 4).

With (u, $\widehat{q}$) pairs at hand, we proceed to align the constituents of u and $\widehat{q}$. The alignment gives us a chunking of u into phrases that map to semantic items in $\widehat{q}$. Alignment is driven by lexicons Lex_e and Lex_r (see Fig. 5), but faces inherent ambiguity, either from truly ambiguous phrases or from noise in the automatically constructed lexicons. We model the resolution of this ambiguity as a constrained optimisation and use Integer Linear Programming (ILP) to address it.

Now we build a bipartite graph with Ph, the set of all phrases from u, on one side and $S_{\widehat{q}}$, the set of semantic items in $\widehat{q}$, on the other. $Ph = ph_1$, ph_2, … is generated by taking all subsequences of tokens in u. We add an edge between each $ph_i \in Ph$ and $s_j \in S_{\widehat{q}}$ where $(ph_i \rightarrow s_j) \in Lex_e \cup Lex_r$ with a weight w_{ij} from the lexicon. Now, for semantic item s_j, E_j, C_j and P_j are 0/1 constants indicating whether s_j is an entity, type, or predicate, respectively. X_{ij} is a 0/1 decision variable whose value is determined by the solution of the ILP. The edge connecting ph_i to s_j in the bipartite graph is retained if $X_{ij} = 1$. Given a set of types connected to a variable v from which we want to pick at most one, this set of types is $S(v) = c_1, c_2$, … and the set of phrases that can map to types in $S(v)$ is $Ph(v)$. Finally, to solve the ILP problem, we are using IBM ILOG CPLEX Optimizer,[10] but we can integrate other solvers programmatically because the system is solver-agnostic.

4.1.3. Template generalisation and system operation

We generalise aligning utterance-query pairs obtained from an alignment process. On the utterance side, we take the utterance u represented using its dependency parse tree and restrict it to the smallest connected subgraph that contains the tokens of all phrases participating in m. To create a template from this subgraph, we turn the nodes participating in m into placeholders by removing their text and keeping the POS tags and semantic alignment annotations (*ent, type, pred*). We use universal POS tags for stronger generalisation power. We replace compound nouns with a noun token that can be used to match compound nouns at testing time to ensure generalisation. At testing time, our templates allow for robust chunking of an incoming question into phrases corresponding to entities (i.e. as named entity recognisers), predicates (i.e. as relation extractors) and types (i.e. as noun phrase chunkers). For NER, we show that using our templates at testing time gives superior results when compared to using an off-the-shelf NER system. On the query side, we take the query and remove the concrete labels of edges (predicates) and nodes (entities and types) participating in m, keeping the semantic alignment annotations. We use the number of utterance-query pairs which generate a template as a signal in query ranking.

When a user presents a new question, $\overline{u}$, in the online phase, a comparison is made against all models in the model repository. First, we determine the dependency parse tree of utterance $\overline{u}$, with its part-of-speech tags. A match to a template ($\overline{u}$, q_t, m_t) exists if there is an isomorphic subgraph of the dependency parse tree of utterance $\overline{u}$ to u_t. For each matching utterance template (usually several), we instantiate the corresponding query template q_t based on the alignment m_t and the lexicon $Lex_e \cup Lex_r$.

4.2. System integration

The proposed KBQA applied to bio-databases reuses two open-source tools, avoiding the development of new components with similar goals. Therefore, we adopted the MONTRA Framework to integrate our tool as a plugin and the SCALEUS-FD to serve as an ontology repository. Fig. 6 represents an overview of the architecture of our proposal. Some of the components of MONTRA Framework and SCALEUS-FD were omitted since these would not increase the value of this description.

The BioKBQA consists of some components that are worth describing. The API Connector can receive questions in natural language and subsequently forward them to the Question Processor. This

[10] https://www.ibm.com/analytics/cplex-optimizer.

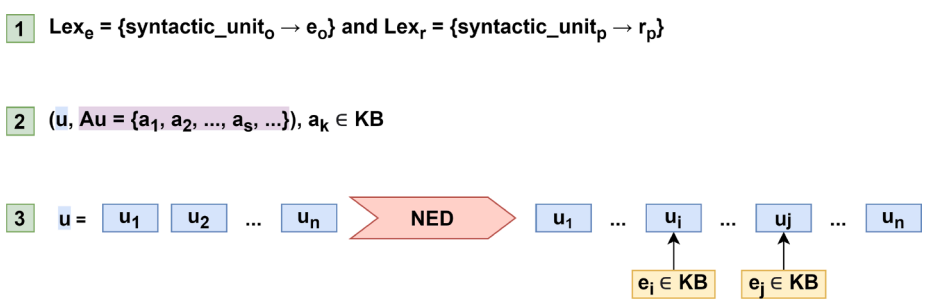

Fig. 2. As a first step, we create the entity and relation lexicons using distant supervision based on rules that map pieces of text to individuals in the KB. Then we consider the pairs of questions in NL and the respective answer entities. In step 3, we split the utterance into its syntactic elements and by a disambiguation process we map surface elements into KB entities.

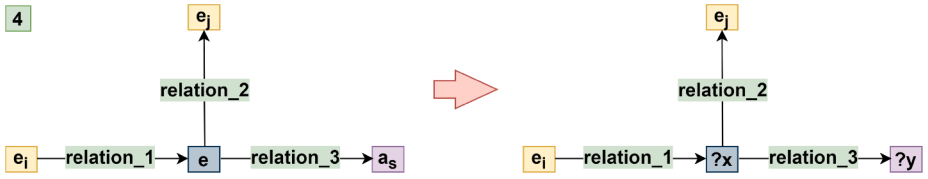

Fig. 3. In this step, we determine the smallest subgraph that links the disambiguated entities to each answer. After that, we replace the new entities found with a variable. Likewise, we replace a_s with a variable.

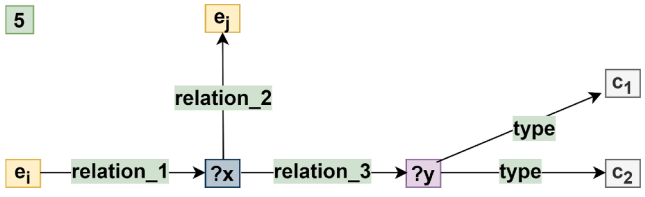

Fig. 4. Looking at the answer a (currently variable ?y), we see which classes it belongs to as an instance. Two classes are shown in the figure, but of course, the number of classes can be different.

component uses the NLP processor to perform the semantic parsing of the query. It also uses Template Management, which serves to match the processed question and the available templates. The SPARQL Processor can extract the desired information from the semantic database. Finally, the responses handled by the Answer Management module are sent to the API Connector, thus ending the processing.

Our proposal aims to help discover datasets of interest based on a research question. This research question is placed on the BioKBQA plugin, integrated into the MONTRA framework. This input set in free-text is converted into SPARQL and sent to MONTRA to obtain the datasets that match this query. MONTRA uses SCALEUS-FD as an ontology repository, which would produce the IRIs of interest for the questions and answers in the data placed on the database catalogue. This would be filtered on MONTRA, retrieving the databases of interest for a question.

4.2.1. Semantic questioning

The question answering (QA) module that we added to SCALEUS-FD allows querying stored semantic data. On the one hand, we can operate traditionally by using SPARQL. This option enables advanced users to exploit a logical query language's power to construct complex queries. Therefore, 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 the linguistic processing tools into the module that would enable us to do semantic parsing. The system processes the information by transforming the NL question into a formal query that the system internally uses to obtain the answers. However, the strength of the solution is the possibility of using templates in the information retrieval process.

To access the module's functionalities, we can use API calls that make it possible to retrieve information through software agents. We have created two endpoints to ask questions using SPARQL or questions in natural language:

- SPARQL endpoint:

GET /api/v1/sparqler/{dataset}/sparql?query={query}&inference={inference}&rules={rules}&format={format} HTTP/1.1.

- NL endpoint:

GET /api/v1/sparqler/{dataset}/nl?query={query}&inference={inference}&rules={rules}&format={format} HTTP/1.1.

A fundamental component of the QA module is the template repository which, together with the parsing unit, allows improved performance in the conversion of complex questions. This repository is fed before putting the tool into production and can be enriched with more templates whenever they are available for use. Fig. 7 shows the offline and online phases of creating and using templates.

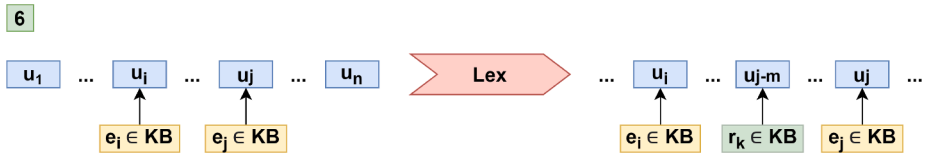

Fig. 5. This step applies the entity and relationship lexicons to find relationships between entities and, possibly, some new entities.

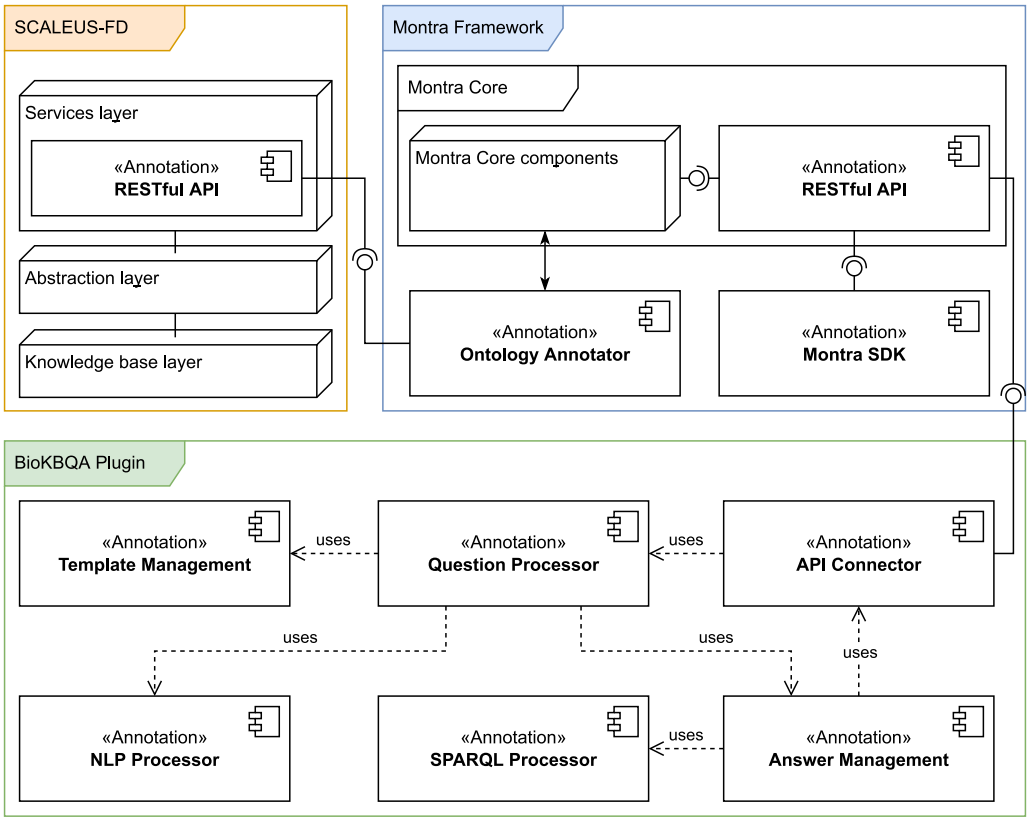

Fig. 6. Component diagram showing the integration of SCALEUS-FD, MONTRA, and the BioKBQA plugin. The MONTRA block is a client of SCALEUS-FD that works as a repository of ontologies and the BioKBQA plugin that allows querying ontologies using natural language.

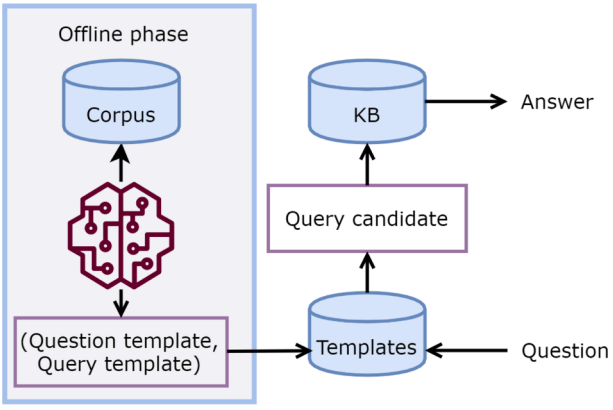

Fig. 7. A high-level view of creating and using templates considering the offline and online phases. A corpus is processed in the offline phase to create pairs of natural language question templates and formal language query templates. Each question is disaggregated in the online stage, with the system running to determine the most suitable model.

5. Results

The different initiatives created to explore one or multiple datasets of patient data usually require some technical background to use the tools designed for filtering and cleaning the data. The use of query builder-like tools is an excellent strategy, but these are typically limited to the data schema and require initial learning by users. Therefore, providing solutions where it is possible to define a question in a free-text format, which will result in a query to be executed in the dataset, may attract users with less technical knowledge.

The proposed methodology can be used by researchers to define simple cohorts over patient datasets, independently of the dataset used. The main overhead of this proposal is the process of defining and mapping the fields and concepts in the database into the ontologies. However, this stage is already performed in some scenarios, with different goals. For instance, there are situations where the ontology is used to enrich the existing knowledge in the data. In other cases, the ontology is used to associate concepts in the data with their standard definition.

5.1. Use case overview

We have identified simple models that provide a good starting point for users to get enough information about datasets of interest. In this way, it is possible to see the validity of using particular datasets in a more profound analysis guided by specialists in the domain. Therefore, we defined three main categories of question answering templates in our methodology: 1) direct questions; 2) questions with exclusive conditions; 3) questions resulting in data aggregation. This approach aims to provide a quick and easy strategy to perform a high-level analysis of each dataset, without having to use sophisticated tools and methodologies. This methodology can be applied in different contexts, as long as we define an ontology to create metadata annotations about the datasets.

The European Medical Information Framework (EMIF)[11] project aimed to improve access to patient-level data from distinct health institutions across Europe, and to carry out multi-cohort studies on different diseases. One of its tracks, the European Medical Information Framework's Alzheimer's disease (EMIF-AD) initiative, aimed to accelerate the discovery and validation of new biomarkers to diagnose

[11] https://www.emif.eu.

Alzheimer's disease in the predementia stage, and to predict the rate of decline. This involved collecting and mapping to an ontology defined for this disease the data of more than 141,050 patients suffering from this disease. The Alzheimer's disease community in this catalogue has currently publicly available information about 65 datasets, with more than 63 still in the addition phase. Each dataset is characterised by more than 480 *meta*-concepts.

5.2. Ontology

Our contribution follows from the work of the EMIF-AD project, where an ontology was constructed to annotate AD data.[12] In parallel, a questionnaire was also made available by this initiative that was used as a skeleton for the construction of the EMIF Catalogue using MONTRA. Our ontology is based on the fields of this MONTRA-loaded questionnaire.[13]

We built an ontology reusing standard medical and biomedical ontologies and vocabularies. We used DCAT to annotate essential information about the repositories described on our platform. We use the DCMI Metadata Terms[14] to annotate bibliographic resources. To report about clinical trials, we use the Ontology for Biomedical Investigations.[15] To describe nuclear radiology entries, we used the RadLex radiology lexicon.[16]

We can present some semantic mappings by way of example. The name of the database is mapped to the DCAT property https://purl.org/dc/terms/title, and the https://purl.org/dc/terms/accessRights term provides access privileges and security status information. To insert a bibliographic reference, we use the term https://purl.org/dc/terms/bibliographicCitation. An exclusion criterion in a clinical trial is annotated with the term https://purl.obolibrary.org/obo/OBI_0500028. A magnetic resonance imaging (MRI) is annotated with the term https://radlex.org/RID/RID10312. Therefore, our ontology follows a hierarchical structure and is subdivided into the following 26 domains: database general information, key publications, data access, inclusion/exclusion criteria, number of subjects, clinical information, dementia and functional rating scales, subjective cognitive impairment, neuropsychiatric scales, quality of life, caregiver burden, health resource utilisation, remote monitoring technologies, cognitive screening tests, neuropsychological tests, physical examination, lifestyle factors, blood collection, cerebrospinal fluid (CSF) collection, urine collection, MRI, positron emission tomography (PET), computerised tomography (CT) scans, single-photon emission computerised tomography (SPECT) scans, electrophysiology, and neuropathology.

FAIRness is guaranteed by the tools we are using. The EMIF Catalogue is a FAIR platform, as demonstrated by Trifan *et al.* [61]. Likewise, SCALEUS-FD is a FAIR tool, as assessed by Pereira *et al.* [63] using the maturity metrics proposed by Wilkinson *et al.* [64].

5.3. Evaluation and error analysis

A researcher interested in analysing Alzheimer's disease datasets could perform a few questions in a free-text format in order to understand the feasibility of the research question before going through the study design, which is time-consuming. For instance, questions that retrieved the number of patients undergoing a specific test during follow-up visits, the number of patients having an exam without taking specific medication, or patients having two or more particular exams. These examples are types of information that fit the three main categories of question-answering templates defined in our methodology.

Considering the lack of an annotated dataset to perform a formal evaluation, we decided to build a set of 30 questions following common research questions, as presented in "The Book of OHDSI" [67], where a broad set of questions is formulated for the creation of cohorts from the consultation of database catalogues. We considered three categories of questions. The first category, C1, consists of questions involving a single concept (Table 1).

In the second category, *C*2, we considered questions with exclusion criteria (Table 2).

Finally, in the third category, C3, we considered multi-concept questions (Table 3).

We manually inspected the database to collect the expected result for each question. Then, we applied our workflow and collected the result for each question, comparing it to the result that was manually obtained. In the cases where a partial matching was achieved we considered it as a wrong answer. After the overall processing, an accuracy of 0.76 was achieved, e.g. 23 questions returned the expected output.

The proposed methodology seems promising in the exploration of semantic datasets. However, we found some limitations. Sometimes the error is because it is impossible to map the relationship between two entities correctly. This is what happens with question 26, "What test is recommended to detect Lewy Body Dementia?". For this question, the disease is registered in the dataset, there are patients entered, and we find mention of tests performed. The problem is that the "is recommended" relationship does not exist. As for question 7, "Which cohorts recorded PET exams?", the "recorded" relation is not successfully disambiguated. So there is no correct triple that can be extracted from the database to get an answer. Likewise, for questions 14, 19 and 20, the system was unable to successfully process the negation expressions "leaving out", "omit", and "unavailable".

Regarding some questions, the process of converting the natural language question and its mapping in a template is not performed correctly. This problem is due to limitations in the NLP processes used to convert surface textual elements into the semantic elements present in the database. For example, we could not define a strategy capable of defining two sets of patients to compare them. An example of a question related to the presented research use case would be question 30, "Between males and females undergoing the CERAD word list exam, which had the higher scores?". The problem is that the system cannot compare two groups of subjects using a global score. This situation refers to the difficulty in mapping order relationships between groups. In this case, it was impossible to return the group (men or women) with the best CERAD word list exam results. There are also problems dealing with multiple logical expressions, as in question 23 where the conjunction "NTB and MRI scans" should be processed before the disjunctive operation "visuoconstruction or (…)".

6. Discussion

The creation of metadata to describe biomedical databases allows researchers to find them in an integrated way. To operationalise semantic operations, a wide range of mature tools can be used to create,

Table 1
Category C1: single concept questions.

Id	Question
1	How many patients performed the neuropsychological examination?
2	How many datasets are there with RM records?
3	Which cohorts contain CT scans?
4	Which cohorts contain computed tomography scans?
5	Which cohorts collected MRI?
6	How many patients are there with MCI?
7	Which cohorts recorded PET exams?
8	Which cohorts collected cerebrospinal fluid patient data?
9	How many databases are there with auditory verbal learning?
10	How many databases are there with NTB data?

[12] https://bioportal.bioontology.org/ontologies/EMIF-AD/?p=summary.
[13] https://github.com/bioinformatics-ua/BioKBQA/blob/master/resources.
[14] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.
[15] https://purl.obolibrary.org/obo/obi.owl.
[16] https://radlex.org/.

Table 2
Category *C2*: questions with exclusion criteria.

Id	Question
11	How many datasets are there without auditory verbal learning test records?
12	How many datasets are there without AVLTs?
13	How many datasets have patients who did not perform the Boston naming test?
14	How many databases are there leaving out PET data?
15	How many datasets are there without visuoconstruction tests?
16	What is the number of datasets without source documentation in Portuguese?
17	What is the number of datasets without CTs?
18	Which databases do not consider Lewy Body Dementia?
19	Which databases omit Lewy Body Dementia?
20	Which databases have neuropathology tests unavailable?

Table 3
Category C3: questions with more than one concept.

Id	Question
21	How many datasets are available with visuoconstruction and neuropsychological test batteries?
22	Which patients performed attention and MRI scan?
23	How many datasets exist with visuoconstruction or with NTB and MRI scans?
24	What are the demographics of patients with CDR information?
25	Which databases have visuoconstruction exams but not neuropsychological test batteries and MRI scans?
26	What test is recommended to detect Lewy Body Dementia?
27	All the patients that performed the Boston naming test and WAIS?
28	How many datasets exist with patients that did not perform the Boston naming test and performed WAIS?
29	Amount of patients that performed a PET exam but did not perform the auditory verbal learning test?
30	Between males and females undergoing the CERAD word list exam, which had the higher scores?

maintain, and store ontologies. Tools like Protégé allow us to build our ontology to capture our knowledge domain. However, we can perform the conversion to semantic data using tools such as SCALEUS-FD, which we can also use as an ontology repository. Furthermore, this tool follows the FAIR principles. Finally, platforms for cataloguing biomedical databases are increasingly common. These catalogues can be built using tools like MONTRA, a solution suitable for building data catalogues for any data domain. Our proposal makes the most of these technologies, augmenting them to overcome their limitations in using natural language so that standard users can find the information they need more easily.

The importance of observational studies for creating new knowledge in areas as diverse as the creation of new drugs or the implementation of new public health policies cannot be overstated. Secondary use of data is naturally only effective if researchers can discover and access biomedical databases suited to their interests. It is typical for initiatives to emerge in the biomedical community attempting to combine the efforts of different actors (patient associations, doctors, researchers) to share data of common interest. This effort translates into creating strategies and tools that can then be used for the benefit of the community. For example, the OHDSI initiative proposes a standard data model and offers tools to query a given database using a query builder directly. But this approach does not allow for discovering other databases and operating in a scenario of interoperability, as is possible with semantic technologies. So, once again, our proposal overcomes these difficulties because it will enable us to search the metadata of database catalogues using a natural language interface for simplicity.

Fig. 8 identifies the various possible steps of an observational study. In the first phase, it is necessary to define precisely the research question for which the study intends to obtain an answer. The next stage establishes the study design and protocol. Here, the researchers define the subjects' inclusion and exclusion criteria and describe the primary and secondary outcomes. It is essential to avoid biases to prevent contaminating this phase with results obtained in later stages of the pipeline,

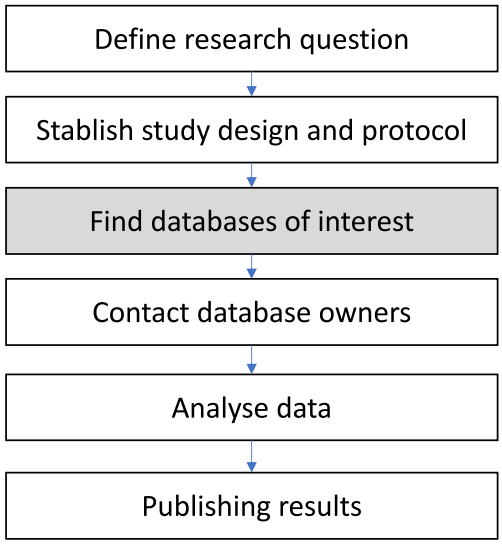

Fig. 8. Typical observational study pipeline, from the research question definition phase to publication of the results.

even if some duly documented recursion is admissible. The step of researching the data of interest is crucial for the study's success. We can use a recommendation system that offers the datasets at this stage [68]. However, this solution is not always flexible and effective as it depends on historical data. Our proposal targets this phase as it allows researchers to locate data efficiently and intuitively. After identifying the relevant databases, the study continues in the following phases: contacting the data owners, defining access policies, analysing data, and publishing results.

Therefore, after specifying the study protocol, a researcher using the proposed system defines the question in natural language that is handled by the BioKBQA plugin, as shown in Fig. 9. This element is integrated into the MONTRA framework to which it forwards the SPARQL query resulting from the processing of the natural language question. MONTRA exchanges messages with the SCALEUS-FD ontology repository, filtering the datasets of interest that return to the user in the last phase. The first message aims to retrieve the entity fields corresponding to the entities present in the translated query. The second message retrieves the IRIs for the answers for each of these fields. This second interaction is required since the data about each database is stored in MONTRA; therefore, SCALEUS-FD cannot filter this in the first interaction.

The BioKBQA plugin is an extension of the system in addition to the two query construction forms available in the system. The first form provides a small set of conjunction-operated fields for building more straightforward questions. A second form, a complete option with all fields with disjunction and conjunction operators, is available but complex to use, which motivated this work.

6.1. System validation

We performed a feedback assessment by a heterogeneous group of users to validate the system's usability. We put together a group of people from different backgrounds, namely bioinformatics researchers, computer science researchers, and medical researchers. None of these users was initially familiar with the system under evaluation. After presenting an instance of EMIF-Catalogue with the new features, we demonstrated how to carry out a search using a form to build the query or using a question in natural language. After this phase, we proceeded in two stages. In the first phase, a set of questions was distributed so that users could use and evaluate each of the data search methodologies. In the second phase, users could experiment freely by asking and building their questions. In the end, a questionnaire was distributed to collect the

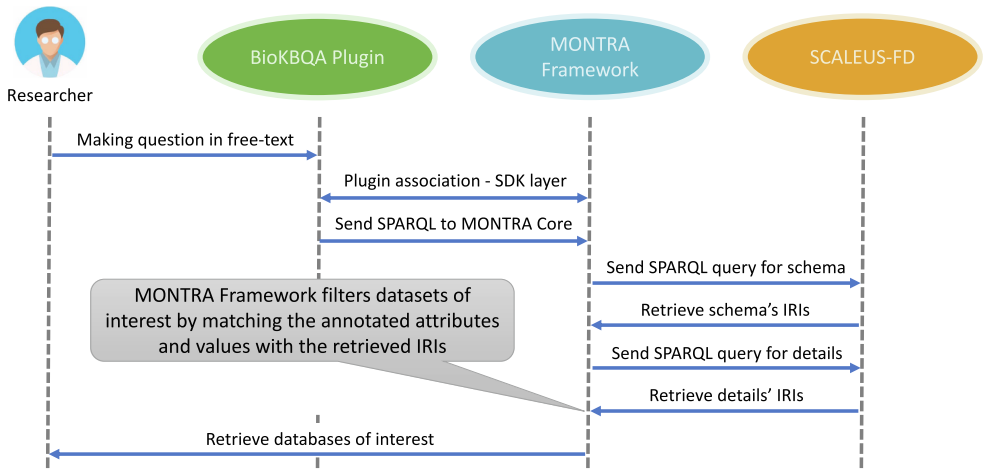

Fig. 9. Interaction diagram showing the interactions between the different systems involved in answering a question asked by a researcher using natural language.

participants' impressions.

A set of six questions similar to "How many patients are there with MCI?" and "What are the demographics of patients with CDR information?" was constructed to test the two opposing alternatives of data search. For each of the questions, we challenged users to use an advanced search form where we can choose from several predefined fields linked by logical operators. Then, we asked users to enter each question into a free text search box. After this round of tests, an extended time was given so that users could create questions on their own using the same two alternatives available: advanced form vs free text input. For this step, we prescribed a set of features and datasets for them to use.

The users' feedback questionnaire was divided into three parts. The first part asks for data related to professional background and mastery of the MONTRA system or similar solutions. In the second part, we evaluated users' satisfaction with the two approaches used and for each test phase. The degree of satisfaction was assessed using a Likert scale, ranging from 1 (i.e., bad) to 5 (i.e., excellent), with the results shown in Table 4. Finally, in the third part, we asked test participants for their ideas and suggestions.

We have drawn some conclusions from the second and third parts of the questionnaire. When given a question to test, it is unanimous that its simple insertion in a text box is easier than using a form. However, when users are asked to construct their questions, the value of using an advanced form increases. This situation indicates that adding an autocomplete mechanism with query suggestions is desirable.

6.2. Future directions

The semantic database that supports the answers to the questions is not always sufficiently complete. Thus, questions well processed by the question-answering module end up not getting a response. This limitation has aroused interest in investigating systems capable of suggesting adjustments to the questions depending on the specific knowledge base. It is also interesting to increase the available data while simultaneously considering unstructured data, such as text. We are then in the domain of hybrid systems, which have also aroused great interest.

Sometimes the created ontologies reveal a limited scope concerning possible questions of interest that researchers need to ask, resulting in lower user adherence because of that data incompleteness. One way to mitigate the incompleteness of ontologies is to find more powerful methods of mining entities and relationships in a text corpus. The idea is to find new entities and relationships, allowing answers to a broader range of questions.

Table 4
Feedback assessment results.

	Bad	Almost fair	Fair	Good	Excellent
Advanced form (phase 1)	0 %	0 %	40 %	60 %	0 %
Free text (phase 1)	0 %	0 %	0 %	20 %	80 %
Advanced form (phase 2)	0 %	0 %	0 %	60 %	40 %
Free text (phase 2)	0 %	0 %	20 %	60 %	20 %

7. Conclusion

Multi-centre studies empower clinical research by extending the research to different populations with similar characteristics. However, finding databases of interest is complex due to the huge number of data partners in the community. Some of these databases are currently characterised in database catalogues, but identifying the right databases using traditional filters is difficult and time-consuming.

The system we propose extends the functionality of biomedical database catalogues to simplify searching for databases. So, in addition to the possibility of using forms to build queries, we can now use an interface that accepts questions in natural language. Our method uses automatically constructed templates and is based on creating an ontology that we use to annotate the descriptors of the databases of interest. Our proposal was implemented using established biomedical tools and was validated considering a catalogue of datasets related to Alzheimer's disease.

Although we applied this system in a catalogue of databases of Alzheimer's disease patients, the technical aspects of this system are not limited to this disease. This strategy can be applied to other, more generic, databases by defining a different ontology.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is funded by national funds through the FCT - Foundation for Science and Technology, Portugal, in the context of the projects UIDB/00127/2020 and DSAIPA/AI/0088/2020. Arnaldo Pereira and João Rafael Almeida are funded by the FCT - Foundation for Science and Technology under the grants PD/BD/142877/2018 and SFRH/BD/147837/2019, respectively. The authors would like to acknowledge with gratitude the support provided by Ana Isabel Morais, GP at USF

Despertar, ACES, Gondomar, Portugal and Guilherme Oliveira, GP at USF Esgueira, Aveiro, Portugal, in the validation of the medical information present in this manuscript.

References

[1] H.G. Cheng, M.R. Phillips, Secondary analysis of existing data: opportunities and implementation, Shanghai Arch. Psychiatry 26 (6) (2014) 371–375, https://doi.org/10.11919/j.issn.1002-0829.214171.

[2] E. Kolker, E. Stewart, V. Ozdemir, Opportunities and challenges for the life sciences community, OMICS: a Journal of, Integr. Biol. 16 (3) (2012) 138–147, https://doi.org/10.1089/omi.2011.0152.

[3] J.C. Wallis, E. Rolando, 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 8 (7) (2013) 1–17, https://doi.org/10.1371/journal.pone.0067332.

[4] J. R, Almeida, O, Fajarda, A, Pereira, J. L, Oliveira, Strategies to access patient clinical data from distributed databases, in: Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies - HEALTHINF, 2019 466–473. doi:10.5220/0007576104660473.

[5] G, Hripcsak, J. D, Duke, N. H, Shah, C. G, Reich, V, Huser, M. J, Schuemie, M. A, Suchard, R. W, Park, I. C. K, Wong, P. R, Rijnbeek, J. v. d, Lei, N, Pratt, G. N, Noŕen, Y.-C, Li, P. E, Stang, D, Madigan, P. B, Ryan, Observational Health Data Sciences and Informatics (OHDSI): Opportunities for observational researchers, Studies in Health Technology and Informatics 216 (2015) 574–578. doi:10.3233/978-1-61499-564-7-574.

[6] L.P. Francis, J.G. Francis, Data reuse and the problem of group identity, Studies in Law, Polit. Soc. 73 (2017) 141–164, https://doi.org/10.1108/S1059-433720170000073004.

[7] K. Höffner, S. Walter, E. Marx, R. Usbeck, J. Lehmann, A.-C. Ngonga Ngomo, M. Sabou, Survey on challenges of question answering in the semantic web, Semantic Web 8 (6) (2017) 895–920.

[8] S. Ferré, E. Hyvönen, Sparklis: An expressive query builder for SPARQL endpoints with guidance in natural language, Semantic Web 8 (3) (2016) 405–418.

[9] A. Pereira, A. Trifan, R.P. Lopes, J.L. Oliveira, Systematic review of question answering over knowledge bases, IET Softw. 16 (1) (2022) 1–13, https://doi.org/10.1049/sfw2.12028.

[10] T. Hamon, N. Grabar, F. Mougin, C. Unger, A.-C. Ngonga Ngomo, P. Cimiano, S. Auer, G. Paliouras, C. Unger, A.-C. Ngonga Ngomo, P. Cimiano, S. Auer, G. Paliouras, Querying biomedical linked data with natural language questions, Semantic Web 8 (4) (2017) 581–599.

[11] B Kern, Mathiak. Are there any differences in data set retrieval compared to well-known literature retrieval?, in: Research and Advanced Technology for Digital Libraries, 2015, pp. 197–208, 978-3-319-24592-8_15.

[12] E, Kacprzak, L. M, Koesten, L.-D, Ibáñez, E, Simperl, J, Tennison, A query log analysis of dataset search, in: Web Engineering, 2017, pp. 429–436. doi:10.1007/978-3-319-60131-1_29.

[13] S. Goel, A. Broder, E. Gabrilovich, B. Pang, Anatomy of the long tail: ordinary people with extraordinary tastes, in, in: Proceedings of the Third ACM Int. Conference on Web Search and Data Mining, 2010, pp. 201–210, https://doi.org/10.1145/1718487.1718513.

[14] D, Brickley, M, Burgess, N, Noy, Google Dataset Search: building a search engine for datasets in an open Web ecosystem, in: Proceedings of the The World Wide Web Conference (WWW), 2019, p. 1365–1375. doi: 10.1145/3308558.3313685.

[15] 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, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M. E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The fair guiding principles for scientific data management and stewardship, Sci. Data 3 (1) (2016), https://doi.org/10.1038/sdata.2016.18.

[16] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L.O.B. da Silva Santos, M. D. Wilkinson, Cloudy, increasingly fair; revisiting the fair data guiding principles for the European open science cloud, Inf. Serv. Use 37 (1) (2017) 49–56, https://doi.org/10.3233/ISU-170824.

[17] P. McQuilton, A. Gonzalez-Beltran, P. Rocca-Serra, M. Thurston, A. Lister, E. Maguire, S.-A. Sansone, BioSharing: curated and crowd-sourced meta- data standards, databases and data policies in the life sciences, Database (2016) 1–8, https://doi.org/10.1093/database/baw075.

[18] Y. Yamamoto, A. Yamaguchi, A. Splendiani, YummyData: providing high-quality open life science data, Database (2018) 1–12, https://doi.org/10.1093/database/bay022.

[19] P. Groth, A. Loizou, A.J. Gray, C. Goble, L. Harland, S. Pettifer, API-centric Linked Data integration: the Open PHACTS Discovery Platform case study, J. Web Semantics 29 (2014) 12–18, https://doi.org/10.2139/ssrn. 3199140.

[20] S.-A. Sansone, A. Gonzalez-Beltran, P. Rocca-Serra, G. Alter, J.S. Grethe, H. Xu, I. M. Fore, J. Lyle, A.E. Gururaj, X. Chen, H.-E. Kim, N. Zong, Y. Li, R. Liu, I.B. Ozyurt, L. Ohno-Machado, Dats, the data tag suite to enable discoverability of datasets, Sci. Data 4 (1) (2017) 1–8, https://doi.org/10.1038/sdata.2017.59.

[21] G. De Moor, M. Sundgren, D. Kalra, A. Schmidt, M. Dugas, B. Claerhout, T. Karakoyun, C. Ohmann, P.-Y. Lastic, N. Ammour, R. Kush, D. Dupont, M. Cuggia, C. Daniel, G. Thienpont, P. Coorevits, Using electronic health records for clinical research: the case of the ehr4cr project, J. Biomed. Inform. 53 (2015) 162–173, https://doi.org/10.1016/j.jbi.2014.10. 006.

[22] J.L. Oliveira, A. Trifan, L.A.B. Silva, E.M.I.F. Catalogue, a collaborative platform for sharing and reusing biomedical data, Int. J. Med. Inf. 126 (2019) 35–45, https://doi.org/10.1016/j.ijmedinf.2019.02. 006.

[23] J. Fan, A. Kalyanpur, D.C. Gondek, D.A. Ferrucci, Automatic knowledge extraction from documents, IBM J. Res. Dev. 56 (3.4) (2012) 1–10, https://doi.org/10.1147/JRD.2012.2186519.

[24] H. Paulheim, Knowledge graph refinement: a survey of approaches and evaluation methods, Semantic Web 8 (3) (2017) 489–508, https://doi.org/10.3233/SW-160218.

[25] G. Schreiber, Y. Raimond, RDF 1.1 Primer, available: W3C Working Group Note (2014) https://www.w3.org/TR/rdf11-primer/.

[26] D.S. Wishart, C. Knox, A.C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, DrugBank: a comprehensive resource for in silico drug discovery and exploration, Nucleic Acids Res. 34 (suppl 1) (2006) D668–D672, https://doi.org/10.1093/nar/gkj067.

[27] T.U. Consortium, The universal protein resource (UniProt), Nucleic Acids Res. 36 (suppl 1) (2007) D190–D195, https://doi.org/10.1093/nar/gki070.

[28] W.N. Borst, Construction of engineering ontologies for knowledge sharing and reuse, University of Twente, 1997. Ph.D. thesis,.

[29] S, Köhler, N. A, Vasilevsky, M, Engelstad, E, Foster, J, McMurry, S, Aymé, G, Baynam, S. M, Bello, C. F, Boerkoel, K. M, Boycott, M, Brudno, O. J, Buske, P. F, Chinnery, V, Cipriani, L. E, Connell, H. J, Dawkins, L. E, DeMare, A. D, Devereau, B, de Vries, H. V, Firth, K, Freson, D, Greene, A, Hamosh, I, Helbig, C, Hum, J. A, Jahn, R, James, R, Krause, S. J. F, Laulederkind, H, Lochmuller, G. J, Lyon, S, Ogishima, A, Olry, W. H, Ouwehand, N, Pontikos, A, Rath, F, Schaefer, R. H, Scott, M, , Segal, P. I, Sergouniotis, R, Sever, C. L, Smith, V, Straub, R, Thompson, C, Turner, E, Turro, M. W, Veltman, T, Vulliamy, J, Yu, J, von Ziegenweidt, A, Zankl, S, Zücbner, T, Zemojtel, J. O, Jacobsen, T, Groza, D, Smedley, C. J, Mungall, M, Haendel, P. N, Robinson, The human phenotype ontology in 2017, Nucleic Acids Research 45 (D1) (2016) D865–D876. doi:10.1093/ nar/gkw1039.

[30] S. Weinreich, R. Mangon, J. Sikkens, M. Teeuw, M. Cornel, Orphanet: A European database for rare diseases, Ned. Tijdschr. Geneeskd. 152 (9) (2008) 518–519.

[31] The Gene Ontology Consortium, Expansion of the gene ontology knowledgebase and resources, Nucleic Acids Res. 45 (D1) (2016) D331–D338, https://doi.org/10.1093/nar/gkw1108.

[32] R, Drysdale, C. E, Cook, R, Petryszak, V, Baillie-Gerritsen, M, Bar- low, E, Gasteiger, F, Gruhl, J, Haas, J, Lanfear, R, Lopez, N, Redaschi, H, Stockinger, D, Teixeira, A, Venkatesan, E. C. D. R, Forum, N, Blomberg, C, Durinx, J, McEntyre, The ELIXIR Core Data Resources: fundamental infrastructure for the life sciences, Bioinformatics 36 (8) (2020) 2636–2642. doi:10.1093/bioinformatics/btz959.

[33] P.L. Whetzel, N.F. Noy, N.H. Shah, P.R. Alexander, C. Nyulas, T. Tu- dorache, 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 Res. 39 (suppl 2) (2011) W541–W545, https://doi.org/10.1093/nar/gkr469.

[34] R. Thompson, L. Johnston, D. Taruscio, L. Monaco, C. Béroud, I.G. Gut, M. G. Hansson, P.-B.A.t. Hoen, G.P. Patrinos, H. Dawkins, M. Ensini, K. Zatloukal, D. Koubi, E. Heslop, J.E. Paschall, M. Posada, P.N. Robinson, K. Bushby, H. Lochmuller, RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research, J. Gen. Intern. Med. 29 (3) (2014) 780–787, https://doi.org/10.1007/s11606-014-2908-8.

[35] L. Penev, D. Koureas, Q. Groom, J. Lanfear, D. Agosti, A. Casino, J. Miller, C. Arvanitidis, G. Cochrane, B. Barov, D. Hobern, O. Banki, W. Addink, U. Kõljalg, P. Ruch, K. Copas, P. Mergen, A. Güntsch, L. Benichou, J.B.G. Lopez, Towards interlinked FAIR biodiversity knowledge: the BiCIKL perspective, Biodiversity Information Sci. Standards 5 (2021) 1–3, https://doi.org/10.3897/biss.5.74233.

[36] S, Shekarpour, E, Marx, A.-C, Ngonga Ngomo, S, Auer, SINA: Semantic interpretation of user queries for question answering on interlinked data, Journal of Web Semantics 30 (2015) 39–51. doi:https://doi.org/10. 1016/j. websem.2014.06.002.

[37] E, Marx, A, Valdestilhas, H, Beck, T, Soru, SANTe: A light-weight end-to- end semantic search framework for RDF data, in: The Semantic Web: ESWC 2021 Satellite Events, 2021, pp, 93–97. doi:10.1007/978-3-030-80418-3_ 17.

[38] E, Marx, K, Höffner, S, Shekarpour, A.-C. N, Ngomo, J, Lehmann, S, Auer, Exploring term networks for semantic search over RDF knowledge graphs, in: Proceedings of the 10th International Conference on Metadata and Semantics Research, 2016, pp. 249–261. doi:10.1007/978-3-319-49157-8_22.

[39] H.k. Azad, A. Deepak, A. Azad, LOD search engine: A semantic search over linked data, J. Intell. Inf. Syst. (2021) 1–21, https://doi.org/10.1007/s10844-021-00687-0.

[40] F. Abad-Navarro, C. Martínez-Costa, J.T. Fernández-Breis, Semankey: a semantics-driven approach for querying RDF repositories using keywords, IEEE Access 9 (2021) 91282–91302, https://doi.org/10.1109/ACCESS.2021.3091413.

[41] B. Ojokoh, E. Adebisi, A review of question answering systems, J. Web Eng. 17 (8) (2018) 717–758, https://doi.org/10.13052/jwe1540-9589. 1785.

[42] V, Lopez, P, Tommasi, S, Kotoulas, J, Wu, QuerioDALI: Question answering over dynamic and linked knowledge graphs, in: Proceedings of the International Semantic Web Conference (ISWC), 2016, pp. 363–382. doi:10.1007/978-3-319-46547-0_32.

[43] A. C, Sima, T, Mendes de Farias, M, Anisimova, C, Dessimoz, M, Robinson- Rechavi, E, Zbinden, K, Stockinger, Bio-SODA: enabling natural language question answering over knowledge graphs without training data, in: Proceedings of the

33rd International Conference on Scientific and Statistical Database Management, 2021, p. 61–72. doi:10.1145/3468791.3469119.
[44] S, Ruseti, A, Mirea, T, Rebedea, S, Trausan-Matu, Qanswer – enhanced entity matching for question answering over linked data, in: Proceedings of the Conference and Labs of the Evaluation Forum (CLEF), 2015, pp. 1–12.
[45] P. Yin, N. Duan, B. Kao, J. Bao, M. Zhou, Answering questions with complex semantic constraints on open knowledge bases, in, in: Proceedings of the 24th ACM Int. Conference on Information and Knowledge Manage., 2015, pp. 1301–1310, https://doi.org/10.1145/2806416.2806542.
[46] S. Hakimov, C. Unger, S. Walter, P. Cimiano, Applying semantic parsing to question answering over linked data: addressing the lexical gap, in, Nat. Language Processing and Information Syst. (2015) 103–109, https://doi.org/10.1007/978-3-319-19581-0_8.
[47] W.-t, Yih, M, Richardson, C, Meek, M.-W, Chang, J, Suh, The value of se- mantic parse labeling for knowledge base question answering, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016, pp. 201–206. doi:10.18653/v1/P16-2033.
[48] D, Song, F, Schilder, C, Smiley, C, Brew, T, Zielund, H, Bretz, R, Mar- tin, C, Dale, J, Duprey, T, Miller, J, Harrison, TR Discover: a natural language interface for querying and analyzing interlinked datasets, in: Proceedings of the The Semantic Web (ISWC), 2015, pp. 21–37. doi:10.1007/978-3-319-25010-6_2.
[49] M, Dubey, S, Dasgupta, A, Sharma, K, Höffner, J, Lehmann, AskNow: A framework for natural language query formalization in SPARQL, in: Proceedings of the European Semantic Web Conference (ESWC), 2016, pp. 300–316. doi:10.1007/978-3-319-34129-3_19.
[50] H, Zafar, G, Napolitano, J, Lehmann, Formal query generation for question answering over knowledge bases, in: Proceedings of the European Semantic Web Conference (ESWC), 2018, pp. 714–728. doi:10.1007/ 978-3-319-93417-4_46.
[51] A, Abdelkawi, H, Zafar, M, Maleshkova, J, Lehmann, Complex query augmentation for question answering over knowledge graphs, in: Proceedings of the OTM Confederated International Conferences "On the Move to Meaningful Internet Systems" (OTM), 2019, pp. 571–587. doi:10.1007/ 978-3-030-33246-4_36.
[52] K. Singh, A. Both, A. Sethupat, S. Shekarpour Frankenstein, A Platform enabling reuse of question answering components, in, in: Proceedings of the European Semantic Web Conference (ESWC), 2018, pp. 624–638, https://doi.org/10.1007/978-3-319-93417-4_40.
[53] K. Singh, A.S. Radhakrishna, A. Both, S. Shekarpour, I. Lytra, R. Usbeck, A. Vyas, A. Khikmatullaev, D. Punjani, C. Lange, M.E. Vidal, J. Lehmann, S. Auer, Why reinvent the wheel: Let's build question answering systems together, in: in: Proceedings of the 2018 World Wide Web Conference, 2018, pp. 1247–1256, https://doi.org/10.1145/3178876.3186023.
[54] L, Dong, F, Wei, M, Zhou, K, Xu, Question answering over Freebase with multi-column convolutional neural networks, 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), 2015, pp. 260–269. doi:10.3115/v1/P15-1026.
[55] K, Xu, S, Reddy, Y, Feng, S, Huang, D, Zhao, Question answering on Free- base via relation extraction and textual evidence, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 2326–2336. doi:10.18653/v1/P16-1220.
[56] D, Lukovnikov, A, Fischer, J. Lehmann, 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, 2017, p. 1211–1220. doi:10.1145/3038912.3052675.
[57] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (56) (2014) 1929–1958.
[58] D, Lukovnikov, A, Fischer, J, Lehmann, Pretrained transformers for simple question answering over knowledge graphs, in: Proceedings of the International Semantic Web Conference (ISWC), 2019, pp. 470–486. doi:10.1007/978-3-030-30793-6_27.
[59] D. Luo, J. Su, S. Yu, A bert-based approach with relation-aware attention for knowledge base question answering, in, Int. Joint Conference on Neural Networks (IJCNN) (2020) 1–8, https://doi.org/10.1109/IJCNN48605. 2020.9207186.
[60] A. Panchbhai, T. Soru, E. Marx, Exploring sequence-to-sequence models for SPARQL pattern composition, in, in: Proceedings of the Iberoamerican Knowledge Graphs and Semantic Web Conference (KGSWC), 2020, pp. 158–165, https://doi.org/10.1007/978-3-030-65384-2_12.
[61] A. Trifan, J.L. Oliveira, A FAIR, marketplace for biomedical data custodians and clinical researchers, in 2018, in: IEEE 31st Int. Symposium on Computer-Based Med. Syst. (CBMS), 2018, pp. 188–193, https://doi.org/10.1109/CBMS.2018.00040.
[62] L.B. Silva, A. Trifan, J.L. Oliveira, MONTRA: an agile architecture for data publishing and discovery, Comput. Methods Programs Biomed. 160 (2018) 33–42, https://doi.org/10.1016/j.cmpb.2018.03.024.
[63] A. Pereira, R.P. Lopes, J.L. Oliveira, SCALEUS-FD: a fair data tool for biomedical applications, Biomed Res. Int. (2020), https://doi.org/10.1155/2020/3041498.
[64] M.D. Wilkinson, S.-A. Sansone, E. Schultes, P. Doorn, L.O. Bonino da Silva Santos, M. Dumontier, A design framework and exemplar metrics for FAIRness, Sci. Data 5 (1) (2018) 1–4, https://doi.org/10.1038/sdata.2018.118.
[65] J, Daiber, M, Jakob, C, Hokamp, P. N, Mendes, Improving efficiency and accuracy in multilingual entity extraction, in: Proceedings of the 9th International Conference on Semantic Systems (I-Semantics), 2013, p. 121–124. doi:10.1145/2506182.2506198.
[66] W.-t, Yih, M.-W, Chang, X, He, 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), 2015, pp. 1321–1331. doi:10.3115/v1/P15-1128.
[67] OHDSI, The book of OHDSI (2022). URL https://ohdsi.github.io/TheBookOfOhdsi/.
[68] J.R. Almeida, E. Monteiro, L.B. Silva, A.P. Sierra, J.L. Oliveira, A recommender system to help discovering cohorts in rare diseases, in,IEEE 33rd, Int. Symposium on Computer-Based Med. Syst.(CBMS) IEEE (2020) 25–28, https://doi.org/10.1109/CBMS49503.2020.00012.