

**Arnaldo António
Pinto Pereira**

Consulta e Visualização de Dados Semânticos

Querying and Visualisation of Semantic Data

**Programa de Doutoramento em Informática
das Universidades do Minho, Aveiro e Porto**



Universidade do Minho



**Arnaldo António
Pinto Pereira**

Consulta e Visualização de Dados Semânticos

Querying and Visualisation of Semantic Data

Tese apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Informática, realizada sob a orientação científica do Doutor José Luís Guimarães Oliveira, Professor Catedrático do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro, e do Doutor Rui Pedro Sanches de Castro Lopes, Professor Coordenador do Departamento de Informática e Comunicações, da Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Bragança.

Programa de Doutoramento em Informática das Universidades do Minho, Aveiro e Porto



Universidade do Minho

Cofinanciado por:



o júri / the jury

presidente / president

Doutora Ana Margarida Corujo Ferreira Lima Ramos
Professora Catedrática da Universidade de Aveiro

vogais / examiners committee

Doutor José Luis Guimarães Oliveira
Professor Catedrático da Universidade de Aveiro (Orientador)

Doutor Francisco José Moreira Couto
Professor Associado com Agregação da Universidade de Lisboa

Doutora Maria Beatriz Alves de Sousa Santos
Professora Associada com Agregação da Universidade de Aveiro

Doutor Gabriel de Sousa Torcato David
Professor Associado da Universidade do Porto

Doutor Enrique Fernández-Blanco
Professor Associado da Universidad da Coruña

agradecimentos

Agradeço aos meus orientadores, Professor José Luís Oliveira e Professor Rui Pedro Lopes, pela valiosa orientação e conselhos. Agradeço ao Instituto de Engenharia Eletrónica e Telemática de Aveiro (IEETA) as excelentes condições para a realização do meu trabalho. Agradeço a todos os meus colegas, com quem tive inúmeras conversas frutíferas, pelo seu apoio e amizade. Em especial, estou grato ao meu bom amigo e colega João Almeida pelo seu forte apoio e encorajamento. Agradeço à minha família, por tudo. Por fim, agradeço à Fundação para a Ciência e a Tecnologia (FCT) que apoiou este trabalho (PD/BD/142877/2018).

acknowledgments

I would like to thank my supervisors, Professor José Luís Oliveira and Professor Rui Pedro Lopes, for their valuable guidance and advice. I gratefully acknowledge the Institute of Electronics and Informatics Engineering of Aveiro (IEETA) for providing excellent conditions to carry out my work. Thank all my colleagues, with whom I had numerous fruitful conversations, for their support and friendship. In particular, I am grateful to my good friend and colleague João Almeida for his strong support and encouragement. Thanks to my family for everything. Finally, I also thank the Fundação para a Ciência e a Tecnologia (FCT) that supported this work (PD/BD/142877/2018).

palavras-chave

Web Semântica, Dados Semânticos, Dados Vinculados, Bases de Conhecimento, Dados FAIR, Interfaces de Linguagem Natural, Pergunta-Resposta, Visualização de Dados.

resumo

As tecnologias semânticas podem descrever dados, mapear e vincular conjuntos de dados distribuídos para uso por pessoas e máquinas. Ao longo dos anos, muitos repositórios de dados semânticos foram disponibilizados na web. No entanto, isso criou novos desafios no que diz respeito à exploração desses recursos de forma eficiente. Normalmente, os serviços de consulta usam linguagens de consulta formais que exigem conhecimento além da experiência do utilizador padrão, o que é crítico na adoção de soluções semânticas. Várias propostas para superar essa dificuldade vêm sugerindo o uso de sistemas pergunta-resposta que fornecem interfaces amigáveis, permitindo entradas em linguagem natural. Por outro lado, processar e integrar os resultados nas formas tabulares usuais não ajuda a entender melhor as informações recuperadas.

Esta tese propõe soluções e métodos para facilitar o acesso e recuperação de informação no contexto de repositórios de dados semânticos. Uma primeira contribuição diz respeito à proposta de uma estratégia de criação e publicação de dados semânticos para diferentes domínios de aplicação, com ênfase em dados biomédicos. Uma segunda contribuição propõe um novo método para aceder aos dados semânticos usando linguagem natural como entrada. Por fim, analisam-se várias possibilidades de visualização de dados semânticos para facilitar sua compreensão e exploração. As propostas foram validadas considerando casos de uso no domínio biomédico usando dados e metadados de pacientes com Alzheimer e pacientes com doença de Huntington.

keywords

Semantic Web, Semantic Data, Linked Data, Knowledge Bases, FAIR Data, Natural Language Interfaces, Question-Answering, Data Visualisation.

abstract

Semantic technologies can describe data, map, and link distributed datasets for people and machines. Over the years, many semantic data repositories have been made available on the web. However, this has created new challenges regarding exploiting these resources efficiently. Usually, querying services use formal query languages requiring knowledge beyond the standard user's expertise, which is critical in adopting semantic solutions. Several proposals to overcome this difficulty have suggested using question-answering systems that provide user-friendly interfaces allowing natural language inputs. On the other hand, processing and integrating the results in the usual tabular forms does not help to understand the retrieved information better.

This thesis proposes solutions and methods to facilitate access and retrieval of information in the context of semantic data repositories. A first contribution concerns the proposal of a strategy for creating and publishing semantic data for different application domains, emphasising biomedical data. A second contribution proposes a new method to access semantic data using natural language as input. Finally, several possibilities for visualising semantic data to facilitate their understanding and exploitation are analysed. The proposals were validated considering use cases in the biomedical domain using data and metadata from patients with Alzheimer's and patients with Huntington's disease.

Table of contents

List of figures	v
List of tables	vii
List of listings	ix
List of abbreviations	xi
1 Introduction	1
1.1 Objectives	3
1.2 Research Methodology	3
1.3 Outcomes	6
1.4 Organisation of the Dissertation	7
2 Semantic Data	9
2.1 The Basics of Semantic Data	10
2.2 Knowledge Representation	15
2.3 Querying Semantic Data	18
2.4 Summary	20
3 Question-Answering over Knowledge Bases	21
3.1 The Basics of KBQA	22
3.2 State-of-the-art of KBQA	27
3.2.1 Semantic Parsing Pipelines	28
3.2.2 KBQA Based on Information Extraction	31
3.3 Challenges and Future Research Directions	32
3.4 Summary	34
4 SCALEUS-FD: A FAIR Data Tool	35
4.1 FAIR Data Principles	35
4.2 Requirements and Building Blocks	37

4.2.1	System Requirements	37
4.2.2	SCALEUS	38
4.2.3	Data and Metadata FAIRness	39
4.3	SCALEUS-FD	40
4.3.1	Architecture of SCALEUS-FD	40
4.3.2	Metadata Hierarchy	41
4.3.3	Implementation	42
4.3.4	Web Services API	43
4.4	Validation	43
4.4.1	FAIR Maturity Assessment	43
4.4.2	Huntington’s Disease Use Case	44
4.5	Summary	46
5	Querying Semantic Data	49
5.1	Contextualisation	50
5.2	Background	53
5.2.1	Discovery of Biomedical Databases	53
5.2.2	Managing Biomedical Data with Semantic Web Technologies	54
5.3	Materials	55
5.3.1	MONTRA Framework	55
5.3.2	SCALEUS-FD	56
5.4	Methods	57
5.4.1	Natural Language Queries over Knowledge Bases	58
5.4.2	System Integration	61
5.5	Results	63
5.5.1	Use Case Overview	64
5.5.2	Ontology	65
5.5.3	Use Case Examples	68
5.5.4	Validation and Error Analysis	68
5.6	Discussion	70
5.6.1	Future Directions	71
5.7	Summary	72
6	Visualisation of Semantic Data	75
6.1	Contextualisation	76
6.2	Background	78
6.2.1	Querying and Visualisation of Semantic Data	78
6.2.2	Interacting with Semantic Data Visualisations	82

6.2.3	Time-evolving Semantic Data	85
6.3	Databases for Observational Health	85
6.4	The EMIF Catalogue Use Case	86
6.4.1	Searching and Visualisation Features	87
6.4.2	Steps for Improved Biomedical Metadata Visualisation	89
6.4.3	Measuring User Behaviour	90
6.5	Ontology-driven Visualisations Scenarios	90
6.5.1	Temporal Knowledge Bases	91
6.5.2	Database-level Visualisations	91
6.5.3	Network-level Visualisations	94
6.5.4	View Refinements	97
6.6	Discussion	99
6.6.1	Impact of Data Visualisations and Interactive Filtering	99
6.6.2	Open Challenges and Future Directions	101
6.7	Summary	102
7	Conclusions and Future Work	105
7.1	Outcomes	105
7.2	Future Work	107
A	Systematic Review Publications	109
B	KBQA Benchmark Datasets Data Samples	115
	References	119

List of figures

1.1	Informal evaluation of the use of semantic data	2
1.2	Research methodology steps	3
2.1	From the web of documents to the web of Linked Data	9
2.2	RDF graph example	12
2.3	Property graph example	12
2.4	DBpedia query example	20
3.1	PRISMA flow diagram and keywords co-occurrence network	21
3.2	Semantic parsing pipelines	24
3.3	Subgraph matching approach	25
3.4	Template-based KBQA general architecture	25
3.5	IE-based KBQA general architecture	26
3.6	Papers by year and architecture	26
4.1	FAIRification process.	37
4.2	SCALEUS-FD architecture and implementation technologies.	41
4.3	SCALEUS-FD metadata.	42
4.4	Spreadsheet integration interface.	45
4.5	QA interface.	46
5.1	Overview of question answering over semantic biomedical data	49
5.2	Query process workflow of common data model-based databases	51
5.3	Query process workflow of semantically annotated databases	52
5.4	General overview of the QA approach	57
5.5	Template-based system starting point	59
5.6	Backbone query creation	59
5.7	Capturing the answer types	60
5.8	Relation disambiguation	60
5.9	Integration of SCALEUS-FD, MONTRA and the BioKBQA plugin	62

5.10	High-level view of template creation	64
5.11	Typical observational study pipeline	71
5.12	Systems involved in answering a question	72
6.1	Screenshot of the YASGUI interface	79
6.2	Screenshot of the PIBAS FedSPARQL interface	80
6.3	Screenshot of the SPEX interface	80
6.4	Screenshot of the SATORI interface	81
6.5	The 15 typologies of network visualisation	83
6.6	EMIF Catalogue database questionnaire form	87
6.7	EMIF Catalogue simple query form	88
6.8	EMIF Catalogue advanced query form	88
6.9	EMIF Catalogue databases comparison view	89
6.10	EMIF Catalogue two-column list selector	89
6.11	UI mockup proposal of a treemap visualisation	92
6.12	UI mockup proposal of graph visualisation	93
6.13	UI mockup proposal of a temporal chart visualisation (entity-level view)	94
6.14	UI mockup proposal of a dendrogram visualisation	95
6.15	UI mockup proposal of map charts visualisation	96
6.16	UI mockup proposal of a heat map visualisation	97
6.17	UI mockup of a temporal chart visualisation (graph-level view)	98

List of tables

1.1	PICO template slot values for building the search query	5
2.1	RDFS constructs	16
2.2	SPARQL endpoints	20
3.1	Question answering over knowledge bases challenges and solutions . . .	33
3.2	Remaining challenges, future work	34
4.1	Semantic namespace	45
5.1	Examples of questions, divided into three main categories	69
A.1	List of publications included in the systematic review of KBQA	109
B.1	KBQA benchmark datasets data samples	115

List of listings

2.1	RDF Turtle example	13
2.2	Shapes graph example	17
2.3	Cypher query example	18
2.4	Excerpt from the SPARQL grammar of the SELECT clause	19

List of abbreviations

ACM	Association for Computing Machinery
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
CC	Creative Commons
CL	Computational Linguistics
CLEF	Conference and Labs of the Evaluation Forum
CNN	Convolutional Neural Network
CSV	Comma-Separated Values
DCAT	Data Catalog Vocabulary
DCMI	Dublin Core Metadata Initiative
EBNF	Extended Backus-Naur Form
EL	Entity Linking
FAIR	Findable, Accessible, Interoperable, and Reusable
FDP	FAIR Data Point
FMD	Fashion Model Directory
FOAF	Friend Of A Friend
GO	Gene Ontology
GRU	Gated Recurrent Unit
HPO	Human Phenotype Ontology
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IEEE	Institute of Electrical and Electronics Engineers
IR	Information Retrieval
IRI	Internationalized Resource Identifier

IT	Information Technology
ITN	Innovative Training Network
JSON	JavaScript Object Notation
JSON-LD	JavaScript Object Notation for Linked Data
KB	Knowledge Base
KBQA	Knowledge Base Question Answering
KG	Knowledge Graph
KOS	Knowledge Organization Systems
LC-QuAD	Large-Scale Complex Question Answering Dataset
LD	Linked Data
LIMSI	Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
LOV	Linked Open Vocabularies
LSTM	Long Short-Term Memory
NCBO	National Center for Biomedical Ontology
NER	Named Entity Recognition
NL	Natural Language
NLIDB	Natural Language Interfaces for Databases
NLP	Natural Language Processing
NNDB	Notable Names Database
OHDSI	Observational Health Data Sciences and Informatics
OO	Object-oriented
OSSE	Open Source Registry for Rare Diseases
OWL	Web Ontology Language
PG	Property Graph
PHI-base	Pathogen-Host Interaction Database
PICO	Population, Intervention, Comparison, Outcomes
POS	Part-of-Speech
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QA	Question Answering
QALD	Question Answering on Linked Data
QG	Query-generation

RDF	Resource Description Framework
RDFa	Resource Description Framework in Attributes
RDFS	Resource Description Framework Schema
REL	Rights Expression Language
REST	Representational State Transfer
RFC	Request for Comments
RL	Relation Linking
RNN	Recurrent Neural Network
SHACL	Shapes Constraint Language
SimpleQ	SimpleQuestions
SKOS	Simple Knowledge Organization System
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
SW	Semantic Web
TE	Textual Entailment
TSV	Tab-Separated Values
Turtle	Terse RDF Triple Language
UniProt	Universal Protein Resource
URI	Uniform Resource Identifier
VANN	Vocabulary for Annotating Vocabulary Descriptions
VQS	Visual Query System
W3C	World Wide Web Consortium
WebQ	WebQuestions
WebQSP	WebQuestionsSP
WLD	Web of Linked Data
WSD	Word Sense Disambiguation
WWW	World Wide Web
XML	Extensible Markup Language
YAGO	Yet Another Great Ontology

Chapter 1

Introduction

Advanced laboratory equipment and increasing digitisation led to large volumes of data and extended life sciences into data-driven sciences (Kolker et al., 2012). The result was a fragmented universe of spreadsheets, databases, non-relational repositories, or just simple raw data dumps, in many cases in the long tail of science and technology, in silos, compromising its reuse (Wallis et al., 2013; Mons et al., 2017). Considering only the clinical and biomedical contexts, one have electronic health record databases (Wade, 2014), patient registries (Sernadela et al., 2017a), omics datasets (Perez-Riverol et al., 2017), medical imaging repositories (Tagare et al., 1997), and virtual biobanks (Jacobs et al., 2018).

Efficient use of secondary data is of paramount importance to improve medical care quality, draw up public health policies, perform pharmacological vigilance, and select patients for clinical trials, to mention a few cases (Schneeweiss and Avorn, 2005). Its use to extract knowledge in the life sciences increased considerably with the surge of data repositories and the digitisation of biobanks (Villanueva et al., 2019). However, this did not immediately translate into a coherent data ecosystem because of heterogeneity, sparsity, and lack of interoperability between distributed data (Golshan et al., 2017).

Researchers continuously struggle to analyse data to answer questions and need solutions to reuse distributed data. They also seek uncomplicated tools for data sharing so that others can benefit, reproduce scientific work, and give credit (Goodman et al., 2014). The use of semantic databases assists in solving data integration and interoperability, allowing the semantic aggregation of information (Berners-Lee et al., 2001; Speicher et al., 2015). They lie at the core of many systems in data-intensive research areas, such as system biology, biopharmaceutics, and translational medicine (Chen et al., 2012). Semantic technologies can describe data and link distributed datasets for people and machines, allowing information search from a single entry point (Paraiso-Medina et al., 2013).

An informal assessment of the popularity of utilising semantic technologies in the biomedical data landscape is to consider the trend of published scientific articles since the introduction of the Semantic Web concept in 2001. PubMed¹ is a resource widely used by medical and biomedical researchers, providing over 33 million life sciences literature records (Sayers et al., 2021). A quick search in this database using keywords related to semantic data and knowledge graphs reveals an exponential interest in semantic technologies, as seen in Figure 1.1(a). One can also observe in Figure 1.1(b) that the creation of life sciences semantic data is a significant portion of the general semantic linked data repositories scenario.

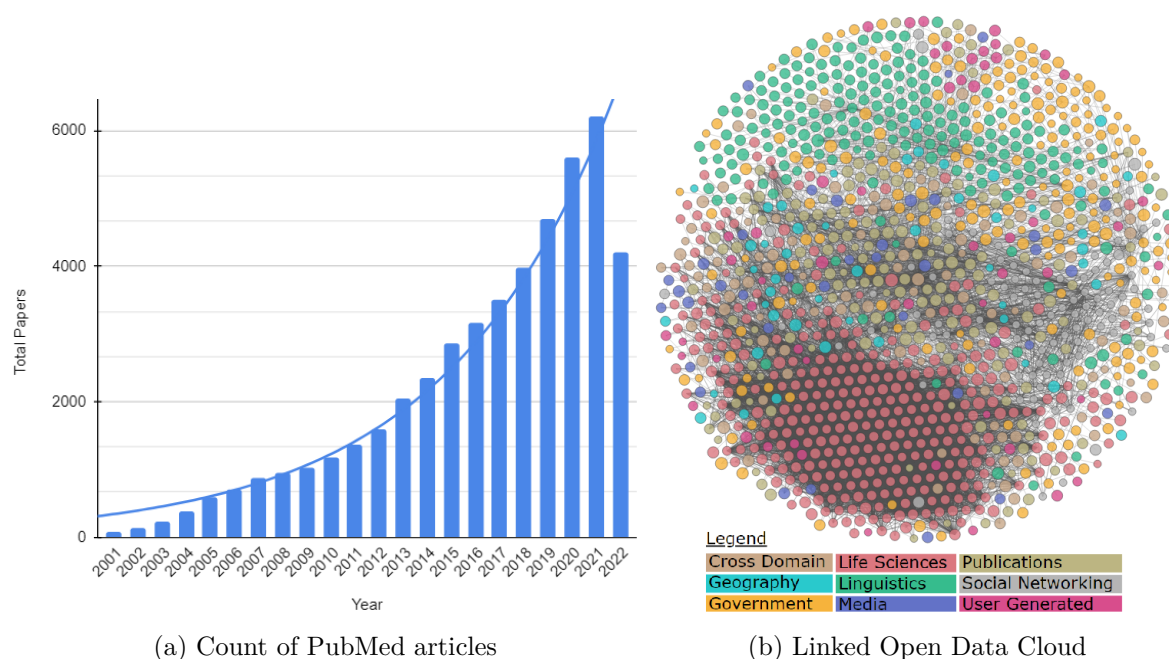


Figure 1.1: Informal evaluation of the use of semantic data. The count of PubMed articles considering semantic data-related keywords is on the left. The figure shows the total number of articles per year in PubMed, retrieved using the search string "semantic web" OR "semantic data" OR "knowledge graph*" OR "ontolog*" and the filter "from 2001 - 2022/7/31". The Linked Open Data Cloud (<https://lod-cloud.net/>), as of May 2020, is on the right. Several life sciences datasets can be spotted as pink bubbles in the lower left of (b).

The explosion of online deployment of semantic databases has raised the question of querying them. On the one hand, there are out-of-the-box query interfaces to input queries in a formal language, but manoeuvring such logical forms is too complex for standard users despite being powerful instruments (Höffner et al., 2017). On the other hand, visual navigation interfaces profiting from the knowledge bases' graph structure primarily facilitate visiting nodes in exploratory walks, but they cannot answer more complex questions (Catarci et al., 1997).

¹<https://pubmed.ncbi.nlm.nih.gov/>

1.1 Objectives

The main objective of this work is to investigate new methodologies that simplify the exploration of semantic data. Therefore, the intention is to answer the following research question:

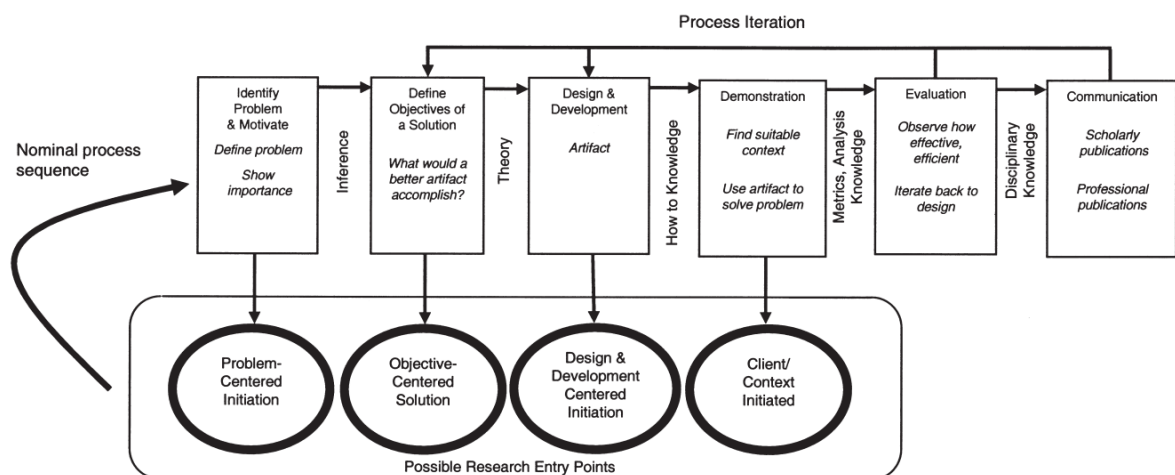
Research Question: Can the use of formal languages be avoided by using natural language to formulate complex questions to obtain answers from biomedical semantic data, and can a varied set of visualisations be used to facilitate exploring this data?

Four sub-goals can be highlighted:

1. Research of state-of-the-art methods and techniques for semantic data querying (Chapter 3).
2. Creation of new techniques for querying semantic data using natural language inputs (Chapter 5).
3. Exploration of semantic data visualisations (Chapter 6).
4. Application of the founded solutions to biomedical use cases (Chapters 4, 5, and 6).

1.2 Research Methodology

The research work was guided by the method proposed by Peffers et al. (2007), which considers an iterative process consisting of six steps, as illustrated in Figure 1.2.



Source: Peffers et al. (2007)

Figure 1.2: Research methodology steps: problem identification and motivation, define the objectives for a solution, design and development, demonstration, evaluation, and communication.

Each of the stages has specific objectives and guided the work carried out, as described below:

1. **Problem identification and motivation** - The purpose of this step is to contextualise and specify the problem. The literature review is the primary tool to understand and integrate previous knowledge as a starting point for creating new knowledge.
2. **Define the objectives for a solution** - Based on the previous step, objectives and requirements are stated for new artefacts that can solve the problem.
3. **Design and development** - This phase relates to the design and development of artefacts embodying the previously theorised proposals.
4. **Demonstration** - The demonstration is performed considering application scenarios or use cases, allowing the instantiation of constructed artefacts to determine if the stated problem is conveniently solved.
5. **Evaluation** - In the evaluation phase, quantitative/qualitative evaluations are carried out.
6. **Communication** - Finally, the produced knowledge is disseminated to relevant audiences through creating written communications and sharing the developed software.

Computer science can benefit from using systematic literature reviews to synthesise the best evidence about state-of-the-art (Kitchenham et al., 2004). A systematic literature review was carried out using a strict methodology to gain in-depth knowledge about the research topic, starting by asking the question: What KBQA methods are there, and what are the solved and unsolved challenges?

From past surveys and overviews, namely Mishra and Jain (2016), Höffner et al. (2017), Ojokoh and Adebisi (2018), Affolter et al. (2019), and Dimitrakis et al. (2020), the keywords shown in Table 1.1 were collected and mapped against the Population, Intervention, Comparison, Outcomes (PICO) structure (Thabane et al., 2009).

Scopus, Web of Science, IEEE Xplore, and the ACM Digital Library were used to find papers, with the search query following the logical form:

$$(\text{Population OR Comparison}) \text{ AND } (\text{Intervention OR Outcomes}) \quad (1.1)$$

Then, studies about KBQA methods or specific KBQA challenges (e.g., modular design, module reusability) were selected. Books, surveys, overviews, tutorials, talks, panel sessions, conference reviews, editorials, abstracts, summaries of workshops or challenges, dissertations, grey literature, and papers not available in English were excluded, as well

Table 1.1: PICO template slot values for building the search query.

PICO Slot	Values
Population	"Knowledge Base*" OR "Knowledge Graph*" OR "Semantic Web" OR "Linked Data*" OR "RDF Data*" OR "data web"
Intervention	Question-Answer* OR "natural language que*" OR "Natural Language Interface"
Comparison	SPARQL OR "Query Graph*"
Outcomes	QALD* OR SimpleQuestions OR WebQuestions OR WebQSP OR LC-QuAD

as those where it was impossible to retrieve the full text. When faced with multiple documents by the same author about the same subject, only those needed to report the core ideas were kept. Articles with unclear, underreported, vague, or inconsistent contributions were also excluded. Furthermore, excluded papers were classified using the criteria listed below to minimise accidental paper rejection further.

- Natural language processing research topics unrelated to KBQA, such as word sense disambiguation (WSD) or textual entailment (TE) recognition.
- Paper on ontologies, taxonomies, or vocabularies. Ontology engineering, ontology learning, or ontology alignment. Knowledge extraction. KB construction and KB completion. KB quality assessment or improvement. Link prediction. Graph embedding, or graph mining. Benchmark dataset.
- Papers about solutions just using formal query languages (e.g., SPARQL) or their extensions (e.g., GeoSPARQL). Query builders, data-semantics-unaware keyword search, or controlled natural language interfaces.
- Papers on question answering over free text, multimedia metadata, or SQL databases. Solution for querying non-semantic data through ontologies. QA on RDF data cubes. Conversational agents. Community question answering. Text mining. Document retrieval and document classification.
- Studies applying existing KBQA solutions without further development.
- Image captioning. Visual question answering. Video question answering. Visual entity linking.
- Mainly about visualisation. Visual query interfaces. Visual query builders.

- Other articles that not comply with the inclusion criteria.

The final selection of papers was guided by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Moher et al., 2009). Regarding the writing of the thesis, the inclusion of an example often obeys the author’s intuition. However, the rule of thumb followed was: choices justified in loco by a bibliographic reference or guiding examples for future fruitful readings. To enable repeatability is available a replication package at <https://osf.io/hxyvw>.

1.3 Outcomes

The dissertation reports on the following outcomes:

- **Systematic review of question-answering over knowledge bases.** Architectural types framing the opponent proposals were considered to meet the challenges posed by implementing KBQA solutions.
- **Use of the Findable, Accessible, Interoperable, and Reusable (FAIR) data principles** to solve the problem of creating, publishing, and accessing semantic data.
- **SCALEUS-FD: a FAIR data tool.** SCALEUS-FD is a FAIR Data software tool for data integration and semantic annotation and enrichment. The core functionalities of the solution follow the SW and LD principles, offering a FAIR REST API for machine-to-machine operations. The source code is publicly available at <https://github.com/bioinformatics-ua/scaleus-fair>.
- **NLP techniques to extract information from structured and unstructured data.** The semantic parsing approach transforms natural language questions into SPARQL by applying various NLP techniques. End-to-end solutions to perform KBQA are based on applying methods to retrieve triples directly from the knowledge base. The proposal to query semantic data using natural language improves these techniques with automatically created templates. The source code is publicly available at <https://bioinformatics-ua.github.io/BioKBQA/>.
- **Exploration of semantic data visualisations.** Different visualisations of semantic data were proposed to support decision-making when choosing biomedical databases.

This document is based on the following papers by the author:

- Rui Antunes, João Figueira Silva, Arnaldo Pereira, Sérgio Matos (2019). “Rule-based and machine learning hybrid system for patient cohort selection.” In: *Pro-*

ceedings of the 12th International Conference on Health Informatics (HEALTH-INF), pp. 59-67. DOI: 10.5220/0007349300590067.

- João Rafael Almeida, Olga Fajarda, Arnaldo Pereira, José Luís Oliveira (2019). “Strategies to access patient clinical data from distributed databases.” In: *Proceedings of the 12th International Conference on Health Informatics (HEALTH-INF)*, pp. 466-473. DOI: 10.5220/0007576104660473.
- Arnaldo Pereira, Rui Pedro Lopes, José Luís Oliveira (2020). “SCALEUS-FD: a FAIR data tool for biomedical applications.” *BioMed Research International*, vol.2020, pp. 1-8. DOI: 10.1155/2020/3041498.
- Arnaldo Pereira, Rui Pedro Lopes, José Luís Oliveira (2021). “Easing the questioning of semantic biomedical data.” In: *Proceedings of the 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 384-388. DOI: 10.1109/CBMS52027.2021.00044.
- Arnaldo Pereira, Alina Trifan, Rui Pedro Lopes, José Luís Oliveira (2022). “Systematic review of question answering over knowledge bases.” *IET Software*, vol. 16(1), pp. 1-13. DOI: 10.1049/sfw2.12028.
- Arnaldo Pereira, João Rafael Almeida, Rui Pedro Lopes, Alejandro Pazos, José Luís Oliveira (2022). “Discovery of biomedical databases through semantic questioning.” *Studies in Health Technology and Informatics*, vol. 294, pp. 585–586. DOI: 10.3233/SHTI220535.
- Arnaldo Pereira, João Rafael Almeida, Rui Pedro Lopes, José Luís Oliveira (2022). “Visualising time-evolving semantic biomedical data.” *Proceedings of the 35th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 264-269. DOI: 10.1109/CBMS55023.2022.00053.
- Arnaldo Pereira, João Rafael Almeida, Rui Pedro Lopes, José Luís Oliveira. “Querying semantic catalogues of biomedical databases.” (Submitted.)
- Arnaldo Pereira, João Rafael Almeida, Rui Pedro Lopes, José Luís Oliveira. “Semantic data visualisation for biomedical database catalogues.” (Submitted.)

1.4 Organisation of the Dissertation

The dissertation has six chapters, besides the introduction, as presented below.

Chapter 2 - Semantic Data. Core concepts about semantic data are presented, starting with the web data model and pointing out connections with graph theory. After

approaching vocabularies and ontologies, the issue of using formal query languages is addressed.

Chapter 3 - Question-Answering over Knowledge Bases. This chapter discusses systems that accept natural language inputs for querying semantic data. The architectures used to build these solutions are visited after the initial concepts presentation. The datasets and benchmarks used to assess them are also discussed. Then, state-of-the-art based on a systematic review of the literature is presented. The final part overviews the remaining challenges and future research directions.

Chapter 4 - SCALEUS-FD: A FAIR Data Tool. This chapter is about a tool that facilitates the creation and deployment of semantic data. The tool follows the Findable, Accessible, Interoperable, and Reusable data principles. An evaluation of the software solution closes the chapter.

Chapter 5 - Querying Semantic Data. This chapter presents a strategy for retrieving semantically annotated biomedical datasets, using an interface built by applying a methodology to transform natural language questions into formal language queries.

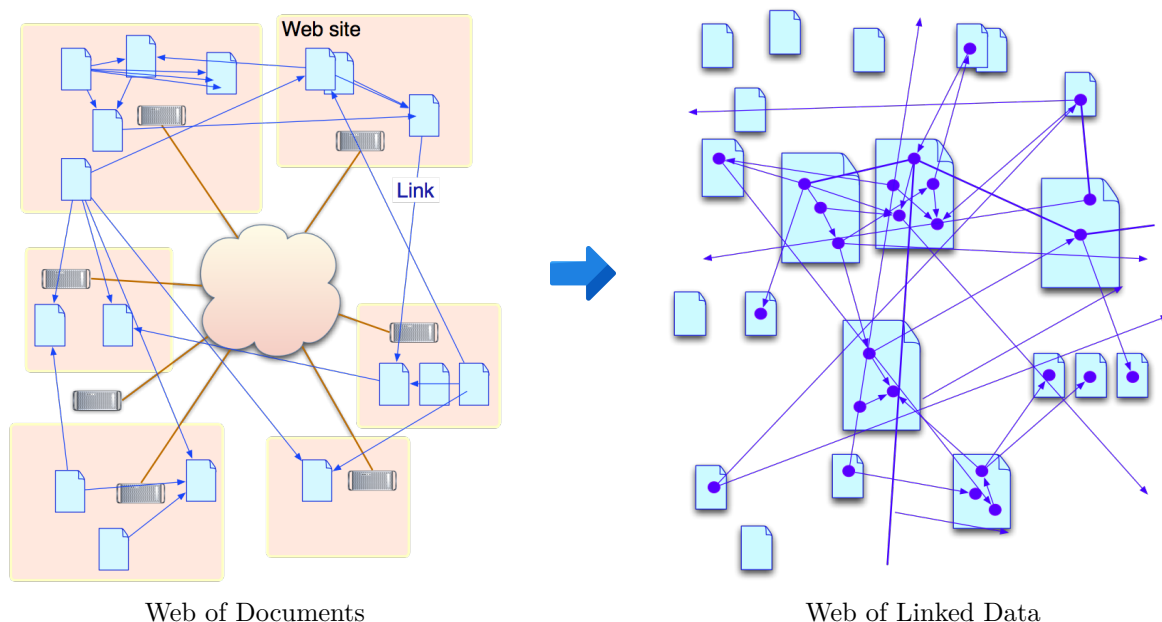
Chapter 6 - Visualisation of Semantic Data. This chapter explores different visualisation and comparison techniques applied to semantic data. This analysis identifies points to improve a catalogue that publishes metadata from multiple health databases, exemplifying the transverse limitations of the most common catalogues. Possible visualisations for semantic information in different health contexts are shown.

Chapter 7 - Conclusions and Future Work. In this final chapter, the research carried out is discussed, summarising the contributions and presenting the limitations and some lines of future work.

Chapter 2

Semantic Data

Semantic technologies make it easier to deal with interoperability and data sharing needs in data-intensive scientific domains. In general, graph data models are suitable abstractions for building knowledge bases. In particular, a special type of graph allowed the creation of a linked data universe by relating entities contained in web documents (see Figure 2.1), with many exciting applications in various domains.



Source: <https://www.w3.org/DesignIssues/Abstractions.html>

Figure 2.1: From the web of documents to the web of Linked Data (WLD). First, one started by linking documents to get information by navigating between them. In the WLD, one connect the entities the documents describe to build knowledge networks for the machine agents to consume.

This chapter establishes semantic data core concepts, starting with the knowledge base definition. Then, a data model proposed for the web is seen and what are vocabularies and ontologies. It is also approached accessing semantic data using formal queries.

2.1 The Basics of Semantic Data

There is no consensus when defining Knowledge Base (KB) or Knowledge Graph (KG) (Ehrlinger and Wöß, 2016; Paulheim, 2017), which will be considered synonymous in this work. A possible definition, based on Mahlmann and Schindelbauer's (2006) formulation of edge labelled multidigraph (or directed labelled multigraph), appeals to graph theory:

A Knowledge Base (or Knowledge Graph) is an edge labelled multidigraph $K = (V, E^)$ that is defined by a node set $V = V_1 \cup V_2$ and a labelled arc set $E^* = \{(v_1, l, v_2) : v_1 \in V_1, v_2 \in V_2, l \in L\}$, l being an element of the label set L . Considering a subset $M \subseteq L$ of arc labels, the M-projection of K is the subgraph K_M composed of all nodes of K and all arcs labelled by the elements of M . The arcs of K_M are called the M-arcs.*

These relatively liberal definitions, intentionally vague about the nature of the elements of sets V , E^* , and L , gets spicier when adding more restrictive conditions to accommodate the World Wide Web Consortium (W3C) RDF (Resource Description Framework) graph standard (Schreiber and Raimond, 2014). Let's start by recalling that an IRI (Internationalized Resource Identifier) is a sequence of characters defined in RFC 3987 (Dürst and Suignard, 2005) that can be used to identify resources (physical and non-physical entities). Literals are associated with a datatype and can optionally be associated with a language tag. Anything in the universe of discourse that is not denoted by an IRI or a literal is called *blank node* (or *bnode* for short). Bnodes represent resources that have not been assigned a specific value (anonymous resources, undetermined objects). They are local in scope, so one must be aware of possible name collisions in operations such as graph union. The following is the definition of RDF graph:

Considering the pairwise disjoint sets \mathcal{I} of IRIs, \mathcal{B} of blank nodes, and \mathcal{L} of literals, an RDF triple (or statement) has the form (s, p, o) , where $s \in \mathcal{I} \cup \mathcal{B}$, $p \in \mathcal{I}$, and $o \in \mathcal{I} \cup \mathcal{B} \cup \mathcal{L}$. s is referred as the subject (the resource being described), p as the predicate (the property, the relationship, the relation), and o as the object (the property value). An RDF graph is a set of RDF triples. A subset of the triples in an RDF graph is a subgraph. A ground RDF graph has no bnodes.

In addition to the definitions, some properties can also be considered (Cyganiak et al., 2014; Hyland et al., 2014):

- RDF graphs are static snapshots of information (atemporality).
- An IRI should never change its intended referent (immutable IRIs, sameness).
- A software agent should not obtain any information about a referenced resource from the sequence of characters composing the IRI reference (opacity of IRIs).
- Literals are constants and never change their value (immutable literals).
- A relation between two resources at one time may not hold at another time (mutable relations).

Cyganiak et al. (2014) also defined entailment, equivalence, and isomorphism. One call *simple interpretations* (or *models*) the concrete arrangements of the world that make an RDF graph true. Considering the RDF graphs K_1 and K_2 , K_1 *entails* K_2 (K_1 is a *semantic extension* of K_2) if every model of K_1 is also a model of K_2 (K_1 truth also makes K_2 true). K_1 and K_2 are *equivalent* if and only if K_1 entails K_2 and K_2 entails K_1 . K_1 and K_2 are *isomorphic* if there is a bijection M between the nodes of K_1 and K_2 , such that: (i) M maps bnodes to bnodes, node literals to node literals, node IRIs to node IRIs, and (ii) $(s, p, o) \in K_1$ if and only if $(M(s), p, M(o)) \in K_2$.

An RDF graph is a KB, with $V_1 = \mathcal{I} \cup \mathcal{B}$, $V_2 = \mathcal{I} \cup \mathcal{B} \cup \mathcal{L}$, and $L = \mathcal{I} \cup \mathcal{B} \cup \mathcal{L}$. An IRI may co-occur as the predicate of one statement and as the subject or object of others. In this case, one node will be considered for all uses as a subject or object and one labelled arc for each appearance as a predicate. Notice that a property is a binary relation. But it is possible to model an n-ary relationship with $n > 2$ using an additional, intermediate node (usually a bnode) or argument lists (Noy and Rector, 2006). Figure 2.2 depicts a graphical representation of an RDF graph in which subjects and objects are the nodes connected by lines (arcs) labelled by the predicates.

Another way currently used by several commercial solutions to implement a KB is to use a Property Graph (PG) (Angles et al., 2017). PGs allow a single node type and not three as in RDF graphs (IRIs, bnodes and literals). A core feature is that arcs and nodes can hold any number of attributes (see Figure 2.3 for an example). Neo4j¹ (Robinson et al., 2015), for instance, is a popular native graph database platform that adopts the property graph model. As there are conversion strategies between property graphs and edge labelled graphs (Das et al., 2014), one can focus only on the latter.

Based on the above arguments, several graph theory concepts can be considered, such as degree, path, distance, neighbourhood, and so on (Angles and Gutierrez, 2008).

¹<https://neo4j.com/>

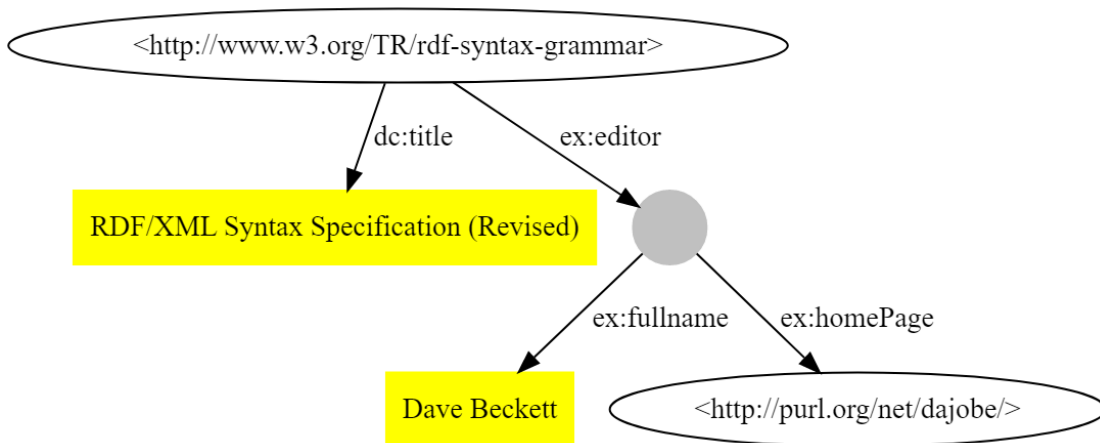


Figure 2.2: The graph presents some relations of the electronic resource identified by the IRI “`http://www.w3.org/TR/rdf-syntax-grammar`” and named “RDF/XML Syntax Specification (Revised),” edited by Dave Beckett. IRI nodes are represented as ovals, all predicate arcs are labelled, and literal nodes are depicted as rectangles. The grey circle represents a bnode (in this case, to model a ternary relation). This figure results from using RDFShape (<https://rdfshape.herokuapp.com/dataInfo>) (Gayo et al., 2018) to process the code in Example 19 of Beckett et al. (2014).

- The node *outdegree* is the number of exiting arcs (equal to the number of triples with the node as the subject). The node *indegree* is the number of entering arcs (the same as the number of triples with the node as the object). For instance, in Figure 2.2, the bnode *indegree* is one, and its *outdegree* is 2. Depending on the problem, these metrics might be related only to a non-empty subset of the labels (relations). More specifically, for an M-projection K_M , the concepts of *M-outdegree* and *M-indegree* can be considered.
- A *path* is a sequence $(e_0, r_1, e_1, \dots, r_n, e_n)$, $n > 0$, of alternating entities and relations. Valid paths have no repeated triples. Depending on the use case, one can consider loops in path construction. The length of a path is equal to its number of arcs. For instance, Hertling et al. (2016) applied these concepts when proposing

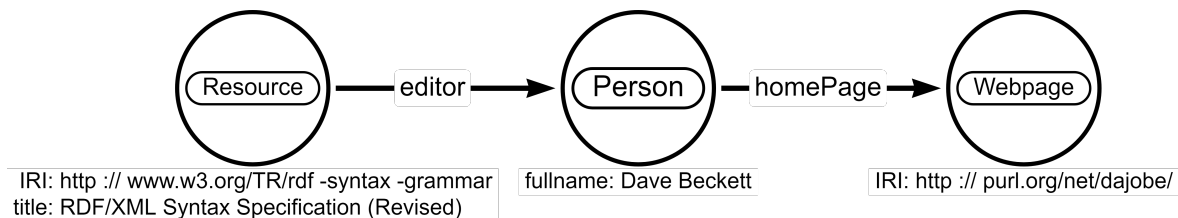


Figure 2.3: The property graph depicted presents the same information as Figure 2.2. The conversion proposed by Angles et al. (2020) was used, which maps into PG nodes the RDF non-literal entities and the bnodes, RDF literal entities into PG attributes, and which maps RDF relations into PG properties. The PG modelling was done using the `arrows.app` (<https://arrows.app/>) tool.

a method to find the k shortest paths between a pair of nodes in an RDF graph. For M-projections, one can speak of *M-paths*.

- Given an RDF graph $K = (V, E^*)$, the M-projection K_M , and a node $u \in V$, the distance from u to v , denoted by $d_M(u, v)$, is the number of M-arcs in the shortest M-path – or infinity if v is not reachable from u (Gubichev et al., 2010). The $N_h(u)$ *h-hop neighbourhood* of u is the set of nodes whose distance from u is less than or equal to h (Khan et al., 2011).

For writing down RDF graphs, concrete syntaxes describe several serialisation formats: N-Triples², Turtle³ (see Listing 2.1 example), TriG⁴, N-Quads⁵, JSON-LD⁶ (JSON-based RDF syntax), RDFa⁷ (for HTML and XML embedding), and RDF/XML⁸ (XML syntax for RDF).

```

1 @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
2 @prefix dc: <http://purl.org/dc/elements/1.1/> .
3 @prefix ex: <http://example.org/stuff/1.0/> .
4
5 <http://www.w3.org/TR/rdf-syntax-grammar>
6   dc:title "RDF/XML Syntax Specification (Revised)" ;
7   ex:editor [
8     ex:fullname "Dave Beckett";
9     ex:homePage <http://purl.org/net/dajobe/>
10  ] .

```

Listing 2.1: RDF Turtle example. A prefix label is associated with an IRI using the `@prefix` directive. A predicate list describes that subject `<http://www.w3.org/TR/rdf-syntax-grammar>` is referenced by several predicates, avoiding writing the full list of triples. An unlabelled blank node is an object in a triple with predicate `ex:editor` and subject in triples with the predicates `ex:fullname` and `ex:homePage` of the nested predicate list.

The true power of the triples is only perceived when considering large datasets. An *RDF store* (or *triplestore*) is a proper database for the storage and retrieval of triples. Some well-known solutions can be listed for quick reference: OpenLink Virtuoso⁹,

²<https://www.w3.org/TR/n-triples/>

³<https://www.w3.org/TR/turtle/>

⁴<https://www.w3.org/TR/trig/>

⁵<https://www.w3.org/TR/n-quads/>

⁶<https://www.w3.org/TR/json-ld11/>

⁷<https://www.w3.org/TR/rdfa-primer/>

⁸<https://www.w3.org/TR/rdf-syntax-grammar/>

⁹<https://virtuoso.openlinksw.com/>

Eclipse RDF4J¹⁰ (formerly known as Sesame), Apache Jena¹¹, and GraphDB¹².

Much more than just creating standardised information about resources, the SW is about data linking. Laying the foundations of the LD paradigm, Berners-Lee (2006) recommended using HTTP URIs for naming things so that people can look up and be provided with helpful information and new links to discover new things:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL).
4. Include links to other URIs. so that they can discover more things.

Some initiatives offering RDF repositories are:

- **DBpedia**¹³ (Lehmann et al., 2015) uses a dedicated open-source extraction framework to extract and turn data from Wikipedia information, such as infoboxes, categories information, geographic coordinates, and external links, into triples.
- **Freebase** (Bollacker et al., 2008) incorporates the Fashion Model Directory (FMD)¹⁴, the Notable Names Database (NNDB)¹⁵, MusicBrainz¹⁶, and Wikipedia data, allowing end-users to do data editions. Presently, only data dumps¹⁷ are available after the initiative termination. Meanwhile, Wikidata integrated Freebase data (Tanon et al., 2016).
- **UniProt** (Universal Protein Resource)¹⁸ (UniProt Consortium, 2020) offers manually curated protein sequences and associated detailed functional annotation.
- **Wikidata**¹⁹ (Vrandečić and Krötzsch, 2014) is a multilingual database that stores facts and the corresponding sources for validity checking purposes.
- **WikiPathways**²⁰ (Martens et al., 2020) is a database of annotated biological pathway models, which are sets of interactions among biological entities, such as proteins and metabolites, regarding a particular context (Hanspers et al., 2021).

¹⁰<http://rdf4j.org/>

¹¹<https://jena.apache.org/>

¹²<http://graphdb.ontotext.com/>

¹³<https://www.dbpedia.org/>

¹⁴<https://www.fashionmodeldirectory.com/>

¹⁵<https://www.nndb.com/>

¹⁶<https://musicbrainz.org/>

¹⁷<https://developers.google.com/freebase>

¹⁸<https://www.uniprot.org/>

¹⁹<https://www.wikidata.org/>

²⁰<https://www.wikipathways.org/>

- **YAGO** (Yet Another Great Ontology)²¹ (Suchanek et al., 2007) is GeoNames²², WordNet²³, and Wikipedia-based, using different heuristics to merge information.

2.2 Knowledge Representation

Defining the web data model to have a way to make assertions about resources is just the starting point for constructing the SW. After that, it is necessary to consider mechanisms that allow users to represent domains of interest by giving semantic meaning to resource IRIs. Adding new concepts is done by semantic extension. This strategy enables the presentation of specific logical-linguistic constructions to make assertions as unique elements for future use with a precise meaning. The most elementary form of knowledge organisation in this context involves defining vocabularies. An *RDF vocabulary* is a set of IRIs establishing entities and relations (jointly referred to as terms) used to describe an area of concern.

RDF Schema (RDFS) is a data-modelling vocabulary providing building blocks (classes and properties) for creating other vocabularies (Brickley and Guha, 2014). Classes are helpful to sort resources into categories. For instance, the `rdf:Property`²⁴ class allows declaring class attributes, while the `rdfs:label` is a property used to provide human-readable resource names. Table 2.1 summarises the primary RDFS modelling constructs.

Standardised vocabularies reuse permits greater efficiency in semanticizing new domains as mapping data to established elements reinforces the interconnection and takes better advantage of the opportunities to infer new knowledge. A dataset uses a vocabulary if a term in that vocabulary appears in the predicate position of a triple or in the object position of a triple whose predicate is `rdf:type` (Schmachtenberg et al., 2014). According to the Linked Open Vocabularies (LOV)²⁵ initiative, the five most commonly used non-outdated vocabularies are as follows:

- **DCMI Metadata Terms**²⁶ describe media (text, images, movies, etc.).
- **Friend of a Friend (FOAF)**²⁷ provides terms to describe people in social networks context.

²¹<https://yago-knowledge.org/>

²²<http://www.geonames.org/>

²³<https://wordnet.princeton.edu/>

²⁴The *namespace IRI* of the IRIs of a vocabulary is a common substring often associated by convention with a short name, the *namespace prefix* (e.g., `rdf`, `rdfs`).

²⁵<https://lov.linkeddata.es/>

²⁶<https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>

²⁷<http://xmlns.com/foaf/spec/>

Table 2.1: RDFS constructs (Schreiber and Raimond, 2014). `rdf:type` is used to state that a resource is a class instance. The `rdfs:subClassOf` and `rdfs:subPropertyOf` properties enable the creation of class and property hierarchies. `rdfs:domain` clarifies the subject of a relation is an instance of a specific class. `rdfs:range` states the values of a property.

Syntactic form	Description
C <code>rdf:type</code> <code>rdfs:Class</code>	Resource C is an RDF class.
P <code>rdf:type</code> <code>rdf:Property</code>	Resource P is an RDF property.
I <code>rdf:type</code> C	Resource I is an instance of class C .
C1 <code>rdfs:subClassOf</code> C2	Class C1 is a subclass of class C2 .
P1 <code>rdfs:subPropertyOf</code> P2	Property P1 is a sub-property of property P2 .
P <code>rdfs:domain</code> C	Domain of property P is class C .
P <code>rdfs:range</code> C	Range of property P is class C .

- **Vocabulary for Annotating Vocabulary Descriptions (VANN)**²⁸.
- **Simple Knowledge Organization System (SKOS)**²⁹ allows describing Knowledge Organization Systems (KOS) like thesauri, classification schemes, taxonomies, etc.
- **Creative Commons Rights Expression Language (CC REL)**³⁰ characterises copyright licenses.

In many situations, it is necessary to encode certain logical aspects, such as formal axiom declaration, to infer knowledge from semantic annotations. The ontology concept covers this need allowing the building of stronger logical formalisation. An ontology formally specifies a shared conceptualisation of a domain (Gruber, 1993; Borst, 1997). When using a formal language like the Web Ontology Language (OWL), it is possible to describe classes, properties, individuals, and data values in a standardised way (Hitzler et al., 2012). The expressiveness of OWL allows for defining not only the vocabulary comprised of terms and relations but also expressing rules for combining terms and relationships, enabling the definition of vocabulary extensions.

Several communities have used ontologies to structure knowledge domains. Regarding life sciences, a couple of examples deserve to be mentioned. The Human Phenotype Ontology (HPO) provides a standardised vocabulary of phenotypic abnormalities encountered in human diseases (Köhler et al., 2016). The Gene Ontology (GO) defines concepts to describe gene function along with three different aspects: molecular func-

²⁸<http://purl.org/vocab/vann/>

²⁹<https://www.w3.org/TR/skos-reference/>

³⁰<https://creativecommons.org/ns>

tion, cellular component, and biological process (Gene Ontology Consortium, 2016). Many more biomedical ontologies and terminologies are available on the NCBO BioPortal³¹.

RDFS and OWL are intended for inference and are unsuitable for OO-type modelling. OO classes have unique attributes (strong typing), while the same RDF relation exists solo and can relate to multiple RDF entities. This flexibility leads to the inability to determine what shape data should take for a specific use. In other words, there is no interface mechanism preventing users from creating meaningless or incomplete data. Shapes Constraint Language (SHACL)³² (Knublauch and Kontokostas, 2017) is a shape language that specifically addresses the need to constrain graph data to a particular shape (see Listing 2.2).

```

1  ex:PersonShape
2    a sh:NodeShape ;
3    sh:targetClass ex:Person ;
4    sh:property [          # _:b1
5      sh:path ex:ssn ;
6      sh:maxCount 1 ;
7      sh:datatype xsd:string ;
8      sh:pattern "^\\d{3}-\\d{2}-\\d{4}$" ;
9    ] ;
10   sh:property [          # _:b2
11     sh:path ex:worksFor ;
12     sh:class ex:Company ;
13     sh:nodeKind sh:IRI ;
14   ] ;
15   sh:closed true ;
16   sh:ignoredProperties ( rdf:type ) .

```

Listing 2.2: Shapes graph example (Knublauch and Kontokostas, 2017).

As seen in Listing 2.2, one can declare the shape that a given RDF graph should have, and it can be verified if a given instance complies with this interface. In the example, `ex:Person` class has two attributes, `ex:ssn` (at most one value) and `ex:worksFor` (unlimited values). The first is a literal of type `xsd:string` while the second is an IRI, instance of the `ex:Company` class.

³¹<https://bioportal.bioontology.org/>

³²<https://www.w3.org/TR/shacl/>

2.3 Querying Semantic Data

Modern semantic query languages like SPARQL (W3C SPARQL Working Group, 2013), Cypher (Francis et al., 2018), and Gremlin (Rodriguez, 2015) are convenient tools for creating, reading, updating, and deleting data from semantic databases. Gremlin is closer to functional programming languages than SQL-like ones, focusing on navigational queries rather than matching patterns. Cypher uses patterns-like building blocks for querying property graphs, following a “pictorial” intuition to encode nodes and edges with arrows between them, as can be seen in Listing 2.3.

```
1 CREATE (:Resource {
2   IRI: "http://www.w3.org/TR/rdf-syntax-grammar", title:
3     "RDF/XML Syntax Specification (Revised)"
4 })-[:editor]->(:Person {
5   fullname: "Dave Beckett"
6 })-[:homePage]->(:Webpage {
7   IRI: "http://purl.org/net/dajobe/"
8 })
```

Listing 2.3: This Cypher query can create the graph presented in Figure 2.3. The “editor” and “homePage” relations connect the “Resource,” “Person,” and “Webpage” nodes. Neo4j Cypher Query Formatter (<https://www.tristanperry.com/cypher-query-formatter/>) was used to format the query exported from arrows.app.

SPARQL 1.1 is the W3C recommendation intended to provide mechanisms for querying and manipulating RDF graphs content. SPARQL has four query forms: **SELECT**, **CONSTRUCT**, **ASK**, and **DESCRIBE**. **SELECT** (see Listing 2.4) returns variables and bindings directly and **CONSTRUCT** returns a single RDF graph. **ASK** returns a boolean indicating whether a query pattern matches or not and **DESCRIBE** returns an RDF graph that describes the resources found (Harris and Seaborne, 2013).


```

1 Query ::= Prologue SelectQuery ValuesClause
2 Prologue ::= (BaseDecl | PrefixDecl)*
3 SelectQuery ::= SelectClause DatasetClause* WhereClause
   SolutionModifier
4 SelectClause ::= 'SELECT' ('DISTINCT' | 'REDUCED')? ((Var |
   ('Expression 'AS' Var')))+ | '*'
5 DatasetClause ::= 'FROM' (DefaultGraphClause |
   NamedGraphClause)
6 WhereClause ::= 'WHERE'? GroupGraphPattern
7 SolutionModifier ::= GroupClause? HavingClause? OrderClause?
   LimitOffsetClauses?

```

Listing 2.4: Excerpt from the SPARQL grammar with the core productions of the **SELECT** clause. **GroupGraphPattern** contains the patterns to be combined with the RDF data. The **SolutionModifier** allows aggregation, grouping, sorting, duplicate removal, or returning only a specific found values window. Extended Backus-Naur Form (EBNF) operators: | (disjunction), ? (zero or one occurrences), * (zero or more occurrences), + (one or more occurrences).

The result from a **SELECT** or an **ASK** query can be serialised as a JSON object or in XML or CSV/TSV formats. In addition, there are also operations to create, update, and remove semantic data. SPARQL also allows graph navigation queries for finding paths between two nodes (Angles et al., 2017).

It is necessary to define a SPARQL endpoint to allow queries to an online semantic repository. A SPARQL endpoint interfaces a knowledge base in a machine-friendly way, using the HTTP protocol to establish a client-server connection. For instance, Figure 2.4 shows a query to the DBpedia SPARQL endpoint³³ and a partial screenshot of the query return: a list of band members and bands.

For a comprehensive list of public SPARQL endpoints, one can go to SPARQLES³⁴ (Vandenbussche et al., 2017), which also has extra information about endpoint sanity checks to assess availability (up/down), performance (cold/warm runs), interoperability (SPARQL 1.0/1.1), and discoverability (Void³⁵ and service descriptions). In this context, availability refers to the ability of the endpoint to respond to a request through the SPARQL protocol. Performance concerns the response time to a SPARQL request over HTTP, and interoperability concerns compliance with the SPARQL 1.1 specification. Finally, the discoverability dimension assesses the degree of self-description of

³³<https://dbpedia.org/sparql>

³⁴<https://sparqls.ai.wu.ac.at/>

³⁵<https://www.w3.org/TR/void/>

The screenshot shows the Virtuoso SPARQL Query Editor interface. On the left, there is a text area for the query. The 'Default Data Set Name (Graph IRI)' is set to 'http://dbpedia.org'. The query text is as follows:

```

PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX dbp: <http://dbpedia.org/resource/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT ?name ?bandname where {
  ?person foaf:name ?name .
  ?band dbo:bandMember ?person .
  ?band dbo:genre dbp:Punk_rock .
  ?band foaf:name ?bandname .
}

```

On the right, a table displays the results of the query. The table has two columns: 'name' and 'bandname'. The results are as follows:

name	bandname
"Lora Logic"@en	"X-Ray Spex"@en
"Steve Diggie"@en	"Buzzcocks"@en
"Tré Cool"@en	"Green Day"@en
"Robert Grey"@en	"Wire"@en
"Robert Gotobed"@en	"Wire"@en
"Billy Zoom"@en	"X"@en
"Erik Sandin"@en	"NOFX"@en
"Colin Newman"@en	"Wire"@en
"Jim DeRogatis"@en	"Vortis"@en
"Pinch"@en	"The Damned"@en
"Baz Warne"@en	"The Stranglers"@en
"Jeff Dean"@en	"The Bomb"@en

Figure 2.4: DBpedia query example using the OpenLink Virtuoso query interface.

the endpoint. As an example, some SPARQL endpoints are listed in Table 2.2.

Table 2.2: SPARQL endpoints.

Database	SPARQL Endpoint
DBpedia	https://dbpedia.org/sparql
UniProt	https://sparql.uniprot.org/
Wikidata	https://query.wikidata.org/
WikiPathways	https://sparql.wikipathways.org/sparql
YAGO	https://yago-knowledge.org/sparql/query

Federated querying over different endpoints is a must-have feature for many LD use cases. SPARQL specification defines the `SERVICE` keyword to allow querying distributed repositories and merging data from various sources (Prud'hommeaux and Buil-Aranda, 2013).

2.4 Summary

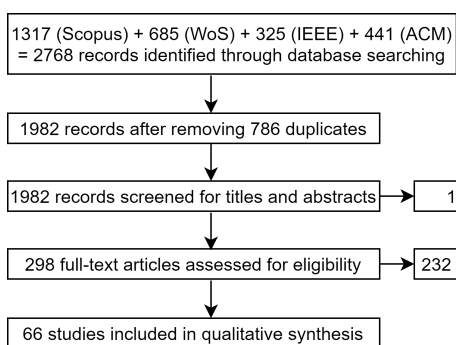
In this chapter, we revisited concepts related to semantic data representation. A tour of the semantic web ecosystem was made, looking at graphs' more general picture, and presenting ontologies to organise knowledge. Finally, using formal languages to query semantic data was introduced.

Chapter 3

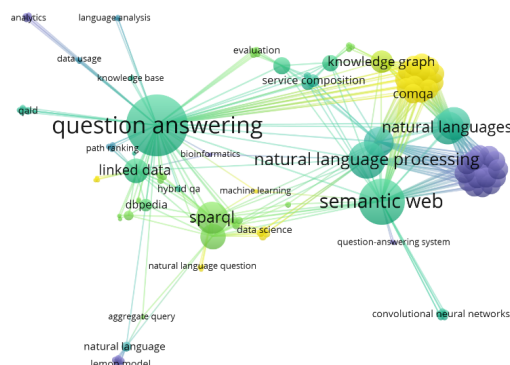
Question-Answering over Knowledge Bases

This chapter presents a systematic review of KBQA methods to identify the main advantages and limitations (Figure 3.1(a) shows the papers selection steps), a topic with the keywords seen in Figure 3.1(b) and at the intersection of Information Retrieval (IR), Computational Linguistics (CL), and SW.

Semantic technologies have enabled the creation of numerous online resources. Still, access to this data is difficult when using formal languages, and a possible help is using Natural Language (NL) interfaces. Question Answering (QA) is about systems that allow users to use NL interfaces to ask questions and receive concise answers. The first



(a) PRISMA flow diagram



(b) Keywords co-occurrence network

Figure 3.1: On the left is the flow diagram of the paper selection process. After searching the bibliographic databases and duplicates removal, there were 1982 records. This number dropped to 298 articles after rejection by screening for titles and abstracts. Finally, after a full-text assessment, 66 studies were eligible for a state-of-the-art review (see Appendix A). On the right is the keyword co-occurrence network (built with VOSviewer - <https://www.vosviewer.com/>) from the literature on KBQA retrieved (see Table A.1). Nodes' size is proportional to the keyword occurrence frequency, and the line thickness represents the co-occurrence intensity between keywords.

QA solution, in the 1960s, intended to answer English questions about baseball games from information saved in a list-structured database (Green et al., 1961). Later, the relational data model gained prominence, and researchers pushed the development of Natural Language Interfaces for Databases (NLIDB). However, just five years after the World Wide Web (WWW) creation, Androutsopoulos et al. (1995) reported the lack of interest in investigating NLIDB. In those days, the focus went to information retrieval techniques to create web search engines using the keyword-based search paradigm. Meanwhile, QA over text was advancing (Hirschman and Gaizauskas, 2001), and the SW vision formulated by Berners-Lee et al. (2001) brought attention to semantic data and KBQA systems.

Search engines started presenting direct answers to some user questions (Guha et al., 2003). Instead of just giving a list of links to documents where the answer is likely to be found, the idea is to satisfy the need for information without further searching and navigation. Questions whose answer is an entity are ideal for this type of approach and using large semantic databases that capture general knowledge has become of great value. In this context, triples extraction to answer questions is priceless and motivates more academic research.

Highlighting the importance of KBQA methods, several researchers using semantic data have been integrating NL interfaces into their systems. To mention just a couple of examples, one can refer to Asiaee et al. (2015), who applied a KBQA solution to parasite immunology, and Hamon et al. (2017), which created a querying platform for linked biomedical data. Other KBQA systems retrieve information from open knowledge databases, such as DBpedia or Wikidata, or proprietary enterprise knowledge graphs, such as Google Knowledge Graph or Bing Satori (Lukovnikov et al., 2017).

Later in this chapter, several proposals framed in different architectural types are presented after recollecting basic concepts. The finale addresses open challenges and future research directions.

3.1 The Basics of KBQA

Considering the nature of the data sources, one can have QA over unstructured data (e.g., text, images), QA over semi-structured data (e.g., graph databases), and QA over structured data (e.g., relational databases). In KBQA systems, the underlying data is semantic data. Hybrid systems are those operating with more than one type of data source. Regarding the scope of data, on the one hand, there are domain-specific solutions when the data schema refers to a particular body of knowledge (e.g., biomedical data) that limits the question types that are accepted. On the other hand, open-domain

systems consider data on generic subjects specified by general ontologies.

Several benchmarks and evaluation campaigns have promoted the advancement of KBQA systems.

- The *Question Answering on Linked Data (QALD)*¹ challenge launched in 2011 is the oldest running campaign, and its ninth edition provided a training dataset with 408 questions in 11 different languages for the open-domain semantic QA over DBpedia task (Usbeck et al., 2018).
- Shortening the QALD dataset size limitations, the *Large-Scale Complex Question Answering Dataset (LC-QuAD)*² provides 30,000 questions with corresponding SPARQL queries for DBpedia and Wikidata (Dubey et al., 2019).
- *Free917*³ is a benchmark with 917 utterances paired with target logical formulas for the Freebase dataset (Cai and Yates, 2013).
- To avoid using logical forms, Berant et al. (2013) created the *WebQuestions (WebQ)*⁴ dataset containing 5810 Freebase question-answer pairs.
- Yih et al. (2016) added SPARQL queries to WebQuestions and created the *WebQuestionsSP (WebQSP)*⁵ benchmark.
- Bordes et al. (2015) achieved, with *SimpleQuestions (SimpleQ)*⁶, a significant scale-up of the numbers with 108,442 questions over Freebase for possible rephrasing in the form *(subject, relationship, ?)*.
- *ComQA*⁷ is a dataset of 11,214 questions collected from WikiAnswers, a community question answering website.
- The *BioASQ*⁸ series of challenges has a task on domain-specific semantic QA on biomedical data to evaluate systems outputting relevant triples and text snippets (Tsatsaronis et al., 2015).

Appendix B allows observing some of the information contained in the different benchmark datasets.

To meet the challenges posed in implementing KBQA solutions, it is important to identify the most common architectures. From the analysis of the papers selected, it was found that they can be classified considering four different architectures. Semantic

¹<https://github.com/ag-sc/QALD/tree/master/9/data>

²<https://github.com/AskNowQA/LC-QuAD2.0/tree/master/dataset>

³<https://nlp.stanford.edu/software/sempr>

⁴<https://worksheets.codalab.org/worksheets/0xba659fe363cb46e7a505c5b6a774dc8a>

⁵<https://www.microsoft.com/en-us/download/details.aspx?id=52763>

⁶<https://github.com/davidgolub/SimpleQA/tree/master/datasets/SimpleQuestions>

⁷<http://qa.mpi-inf.mpg.de/comqa/>

⁸<http://bioasq.org>

parsing pipelines are solutions based on semantic parsing, which uses a pipe and filter style where data flows to generate a formal query from the original input in NL. It is the most straightforward architectural style of KBQA systems and relies on connecting components to form a pipeline, as shown in Figure 3.2.

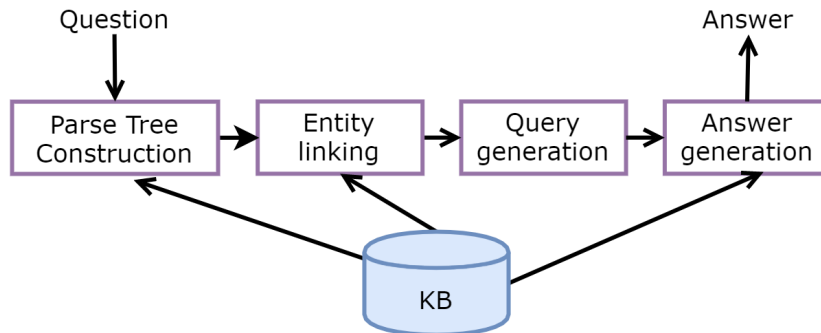


Figure 3.2: General architecture for semantic parsing pipelines. The direction of the arrows denotes the direction of the data flows. Generally speaking, four phases are at play: the division of the question into linguistic units, linking linguistic elements to KB objects, the creation of a formal query, and, finally, generating the answers.

The idea is to apply several data transformations from the question in NL to the logical form or formal query. To achieve that, natural language processing (NLP) techniques such as tokenization, part-of-speech tagging, named entity recognition, dependency parsing, and entity/relation linking are used.

- *Tokenization* is the task of breaking a string of characters into pieces, called tokens, eventually discarding certain characters such as punctuation.
- *Part-of-speech tagging* (also known as *POS tagging*) assigns a part of speech like NOUN or VERB to each input word establishing its grammatical role in the sentence.
- *Named entity recognition (NER)* allows assigning tags referring entities from a lexical resource, like PERSON, LOCATION, or ORGANIZATION, to sets of words.
- *Dependency parsing* determines the grammatical structure and relationships between the words of a sentence.
- *Entity linking (EL)* is the task of assigning a unique KB individual to an entity mentioned in a text.
- *Relation linking (RL)* is the task of assigning a unique KB individual to a relation mentioned in a text.

An alternative way of using semantic parsing is based on the observation that executing a formal query is equivalent to finding a subgraph, as depicted in Figure 3.3.

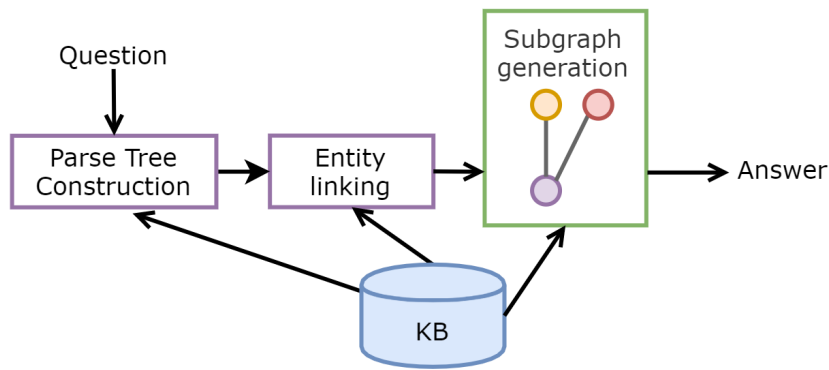


Figure 3.3: Subgraph matching approach.

Systems capable of answering complex questions (e.g., questions that cannot be reduced to a simple triple pattern) require more sophistication than the systems presented so far. A template is a query skeleton with an arbitrary degree of complexity that fits the knowledge base to be questioned and has slots that must be filled with information from entities and relations. The quality of the template-based system depends on the effort put into creating the templates. These systems rely on the manual or automatic creation of a template database assuming an architectural configuration such as that shown in Figure 3.4.

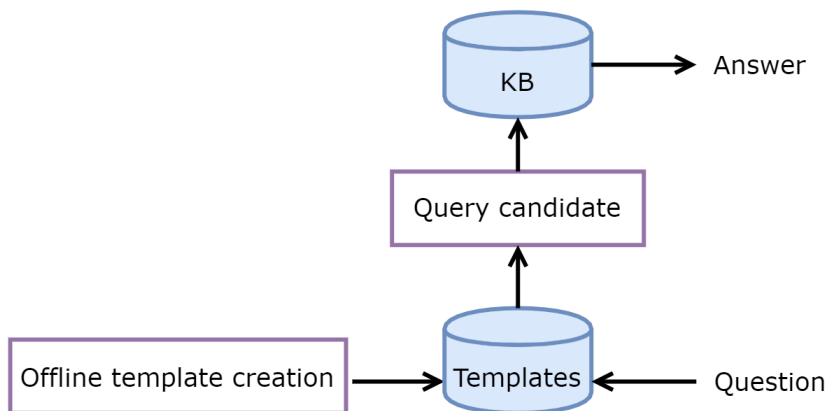


Figure 3.4: Template-based KBQA general architecture.

In the offline phase, it is necessary to create templates. This involves considering pairs of questions and answers used to obtain successively more abstract representations that are used to generate pairs of question-query templates after alignment. The online phase is straightforward: a question is matched with a template to produce a query template, the slots are filled with entities and relations, and the answer is provided by issuing the query candidate.

End-to-end solutions perform sequence-to-sequence translation or apply methods to extract triples directly from the KB. The selection of the final answer is based on the representations of the questions in NL obtained by applying machine learning techniques, as can be seen in Figure 3.5. After extracting the candidate answers from the KB, they are evaluated against a predefined score using a specialized function.

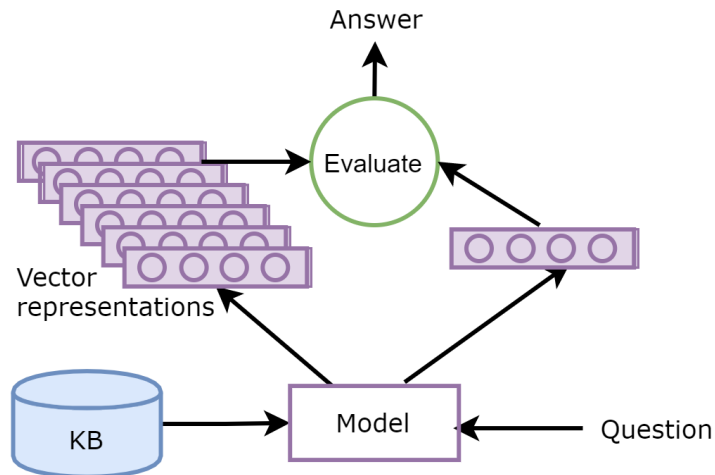


Figure 3.5: IE-based KBQA general architecture.

Figure 3.6 shows the distribution of the selected articles divided by types of architecture and distributed over years. As can be seen, there is a consistent decline in the use of pipeline-based approaches. On the other hand, after an increase in subgraph matching solutions, a slight drop in 2020 is observable. After a boom in 2016, the proposals for information extraction fell to a plateau still higher than the other proposals. Finally, template-based systems fluctuated to an annual maximum of two proposals in 2017 and 2018.

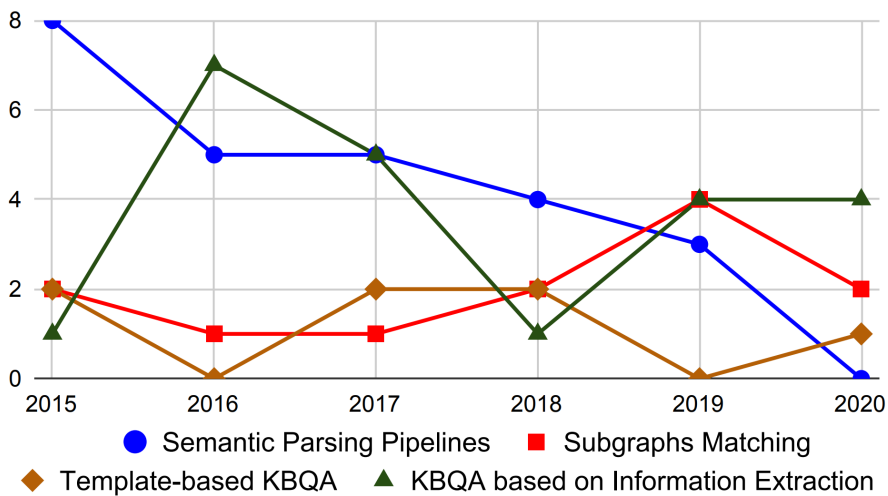


Figure 3.6: Distribution of papers by year and architecture.

Höffner et al. (2017) highlighted significant challenges faced by semantic QA. The lexical gap occurs when the surface forms used in a question are different from those used in the KB. The ambiguity stemming from the fact that the same word can represent various entities is also problematic. Another significant problem is finding answers to questions manoeuvring several units combined in complex queries requiring ordered, aggregated, or filtered outputs. Equally challenging is multilingualism, which concerns two distinct realities that may or may not co-occur. The first involves the problem of using the same interface to ask questions in several NLs, and the second has to do with the possibility of the KB data being multilingual. However, systems relying on languages other than English end up receiving far less attention from the scientific community, limiting the number of available solutions. For instance, very few developers have participated in challenges like QALD with multilingual systems. Some systems try to prevent difficulties by using controlled natural languages (CNLs), which are constructions that restrict in some way the lexicon, syntax, or semantics of the NL from which they start. This review does not focus on multilingualism or the use of CNL interfaces.

3.2 State-of-the-art of KBQA

A solution for retrieving facts from a semantic database is to use semantic search engines based on keywords (Shekarpour et al., 2015). SANTé (Marx et al., 2021) allows the publication, browsing, and querying of arbitrary RDF data. SANTé’s keyword-based search engine relies on building a network of terms using the values of the `rdfs:label`⁹ property, following the formalisation of Marx et al. (2016). Azad et al. (2021) 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 backward search starts from the object to the subjects. Another approach is Semankey (Abad-Navarro et al., 2021) 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 (Ojokoh and Adebisi, 2018).

The solutions for creating natural language interfaces for knowledge bases can be

⁹https://www.w3.org/TR/rdf-schema/#ch_label

divided into two main groups: 1) semantic parsing and 2) information extraction. In the first group, it is performed semantic parsing applying NLP techniques intending to transform the NL question into a formal query that is used to obtain the answers, ending the process. In the second group, solutions can be found to get the answers by extracting information directly from the knowledge base without creating a formal query.

3.2.1 Semantic Parsing Pipelines

Hamon et al. (2017) use a multi-step method to answer the QALD-4 Task 2¹⁰ biomedical interlinked data questions. NL questions go through an annotation process and a linkage phase of surface forms to biomedical entities. In query construction time, fixed rules allow using the previously identified elements to build a SPARQL query. Similarly, the QuerioDALI solution (Lopez et al., 2016) first performs a NER to classify named entities, and then an EL filter binds a unique identity to each entity identified in the previous step. Finally, the system uses fusion and ranking of possible answers.

Ruseti et al. (2015) use DBpedia and Wikipedia to map NL question phrasal constructs to ontology entities. To address the lexical gap, Yin et al. (2015) perform question paraphrasing. Hakimov et al. (2015) consider a combinatorial categorical grammar with handcrafted lexical items and lambda-type calculus expressions to obtain semantic representations. In this way, the input utterances must comply with the grammar. As is naturally emphasized by the authors, performance improves according to the lexicon size. Yih et al. (2016) reached the same conclusion, showing that learning from labelled semantic parsers improves overall performance.

TR Discover (Song et al., 2015) solution uses a grammar that maps first-order logical expressions to SPARQL. Dubey et al. (2016) also propose a grammar but consider an additional normalisation step to create intermediate canonical syntactic forms representing NL questions.

The query-generation (QG) process of a QA pipeline occurs after the entity and relation linking subtasks. Zafar et al. (2018) start with the identified entities and relations and generates walks on the KB by using the adjacent connections within a one-hop distance. Valid walks are the ones containing all the starting entities. Finally, the creation of SPARQL queries occurs after evaluating the candidate walks against the question type. To extend QG to ordinal and filter questions, Abdelkawi et al. (2019) added extra constraints to the list of all possible answers.

Several KBQA-related contributions can be reported as part of the WDAqua Marie

¹⁰<http://qald.aksw.org/index.php?x=task2&q=4>

Skłodowska Curie ITN¹¹ effort to advance the QA field. Both et al. (2016) start from the realisation that QA systems are very complex and usually monolithic to present Qanary¹², a vocabulary-driven methodology to allow decoupling of the different semantic pipeline parts and thus achieve reconfiguration and reuse. First, the Web Annotation Data Model¹³ is used to create a vocabulary covering the common abstractions related to the authors' idea of a QA pipeline. In addition, the input and output of filters are described to achieve interoperability, forcing the components to have the same interface, like in a uniform pipe and filter architecture. Diefenbach et al. (2017b) proposed using timestamps to avoid conflict between Qanary annotations when changing module input descriptions at runtime to allow user feedback. Considering that no vocabulary can describe all existing modules, the burden of creating a new description is to component developers, making it hard for methodology adoption. The problem of adapting the input and output of each module to comply with the shared vocabulary is also burdensome. On the other hand, Diefenbach et al. (2017a) presented a reusable user interface to call the Qanary APIs to make life easier for end-users.

The idea of creating a generic (pipeline) architecture for QA on linked data to foster cooperation among developers is championed by QAestro¹⁴ (Singh et al., 2017), a proposal competing with Qanary that can be used to combine building blocks in tailored systems, allowing a semantic description of both QA components and requirements. Several important subtasks are covered, such as tokenization, POS tagging, NER, EL, dependency parsing, triple generation, data mapping, QG, and answer generation. Question type identification, answer type identification, query ranking, and syntactic parsing are also available.

Embracing the quest for component reuse, Frankenstein¹⁵ (Singh et al., 2018a) is a platform that collects several core components to solve QA tasks and enables the creation of different QA pipelines, more precisely 380 when the paper was published. Highlighting the fact that modern QA systems rely on the flexible integration of many specialized filters, Singh et al. (2018b) suggests that the construction of the pipeline could be considered an optimization problem, where each component could be selected from a set of options for NER and EL, relation extraction and query building. The prediction of the best-performing components facing a new NL question is tackled as a supervised learning problem.

The use of semantic pipelines for KBQA is the oldest and most documented approach

¹¹<https://github.com/WDAqua>

¹²<https://github.com/WDAqua/Qanary>

¹³<https://www.w3.org/TR/annotation-model/>

¹⁴<https://github.com/WDAqua/QAestro>

¹⁵<https://github.com/WDAqua/Frankenstein>

in the literature and is preferred by authors who intend to integrate NL interfaces into their systems quickly. Reinforcing this statement is the existence of frameworks that allow the decoupling of the different components used to filter the data, thus offering greater customization. It is also the easiest way for those who do not want to invest a great deal to develop more technically elaborate solutions, usually with better performance. Each filter can be independently investigated because they are of interest in many other applications, not just in QA. For instance, Shen et al. (2015) surveyed EL issues, techniques, and solutions. Nevertheless, this way of solving the problem seems to be reaching its maturity and more important future developments will almost certainly come from other approaches.

Some proposals depart very little from the classic pipeline, building the query subgraph using a semantic tree, whereas others move away sharply by constructing the subgraph step by step from a starting entity. Hu et al. (2018a) start by finding the semantic tree, and then after extracting the semantic relations, they build a semantic query graph. More elaborately, Yih et al. (2015) propose staged query graph generation, a solution that formulates a query graph by solving a search problem. A general query subgraph is supported by the existing entities in the KB, an existential node not mappable to the KB, and a node for identifying possible aggregation functionality. The solution revolves around creating an inferential chain starting with a root entity node and using legitimate actions to grow a query graph. The first step is to find root candidates by using a lexicon to perform EL over the input query. The next step considers the lexicon again to extract the expected answer. By relating the root entity and the kind of answer, it is possible to create a set of candidate subgraphs constrained by an aggregation function. Finally, a convolutional neural network is used to select the best candidate. For this last classification task, one can use the proposal by Maheshwari et al. (2019), which considers a self-care mechanism that explores the intrinsic structure of subgraphs.

Zheng et al. (2015) started from an initial set of NL questions and formal queries to propose a technique based on studying the similarity of graphs generated from the utterances and SPARQL queries to match the best candidate pairs to form a database with templates. Savenkov and Agichtein (2016) used external text data to explore the central topic of the question and select the best query candidates using a predefined collection of query templates. However, considering a set of manually adjusted templates is necessarily limiting, for instance, when new relations are added to the KB. Literature offers few proposals for this type of system, despite allowing answers to a wide range of questions. Investing in research to create wider lexicons to be used in the production of templates promises the creation of systems with even

higher performance regarding complex questions. However, it seems that the research effort is shifting to end-to-end systems.

3.2.2 KBQA Based on Information Extraction

Several KBQA solutions using some form of a deep neural network have been reported. Dong et al. (2015) introduced a multicolumn convolutional neural network to understand questions from three different aspects, answer path, answer context, and answer type, and learn their distributed representations. Meanwhile, the system enables joint learning of low-dimensional embeddings of entities and relations in the KB. This approach can be expanded and enriched by considering more dimensions to convert into vector representations. Xu et al. (2016b) present a neural network-based relation extractor to retrieve the candidate answers from Freebase and then infer from Wikipedia to validate these answers. More precisely, the process involves dividing the original question into subquestions by applying a set of syntactic patterns. Then, for each subquestion, EL and relation extraction is performed and refined by a joint inference model. After retrieving a set of candidate answers, the final solution is obtained by inference on Wikipedia, searching on the page of the topic entity for evidence about candidate answers.

The model proposed by Lukovnikov et al. (2017) learns to rank subject–predicate pairs to enable the retrieval of relevant facts given a question. The network contains a nested word and character-level question encoder that allows the handling of new and rare words without compromising the exploitation of word-level semantics. This neural network approach generates a single process solution that avoids complex NLP pipeline constructions and error propagation, and it can be retrained or reused for different domains. In scenarios where training data is limited, overfitting compromises network performance. To tackle this problem, instead of using a bidirectional long short-term memory (LSTM) network to create the language representation model, Lukovnikov et al. (2019), Luo et al. (2020a), and Panchbhai et al. (2020) independently evaluated the use of Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), the current most performant solution for NL understanding tasks.

Hao et al. (2017) present a model to represent the questions and their corresponding scores dynamically according to the various candidate answer aspects via the cross-attention mechanism. In addition, they leverage the global knowledge inside the underlying KB, aiming to integrate this information into the representation of the answers. As a result, it could alleviate the out-of-vocabulary problem, which helps the cross-attention model to represent the question more precisely.

Relation detection is essential to extract candidate answer triples. Yu et al. (2017)

use deep residual bidirectional LSTM networks to compare questions and relation names considering different abstraction hierarchies. This relation detector integrates EL for mutual enhancement, similar to the joint inference feature of Xu et al. (2016b).

The creation of models to generate vector representations of features of interest from KB avoids the use of semantic pipelines. As there are multiple architectures of deep neural networks and varied ways of digesting the information to be processed, the literature already reports several possibilities, and many more will appear shortly. LSTMs with attention have great room for further development. On the other hand, transfer learning using pre-trained models is still underrepresented in new system implementations. Finally, the arrival of new and better-performing models allows better results but at computational costs that are not always bearable.

3.3 Challenges and Future Research Directions

Several obstacles have prevented the full adoption of KBQA systems. Table 3.1 presents a summary of the challenges KBQA has faced. The preferred technique for solving simple questions is sequence-to-sequence translators using neural networks. An encoder converts the NL question to a vector representation, and then a decoder outputs a query in a formal language. It is also possible to extract features by processing convolutions.

There are several research proposals on complex questions, starting with systems that propose adding support to another set of SPARQL modifiers. More sophisticated techniques such as the generation of templates or the use of subgraphs are also on the agenda. The information extraction approach using a neural model is also reported. Hybrid systems that use KB data and free text were also found. This technique is also used to mitigate KB incompleteness. The renewed interest in both topics indicates that these challenges are not closed. Entity and relation linking are unsolved issues, although the joint entity and relation linking approach shows promise. Automatic labelling and distant supervision usually help in obtaining more training data.

In general, almost all papers promise to tune their proposals for better performance. However, two major problems remain open, as presented in Table 3.2. Future work to tackle the answers to complex questions revolves around exploring solutions that allow real-time feedback to the system, such as implementing a conversational agent or shifting to reinforcement learning so that new knowledge adds can be continuous. On the other hand, KB incompleteness also limits these systems' usability. Hybrid systems that use free text to address this problem have been explored, but there is still a long way to go. There is a need for more training data and more external knowledge.

Table 3.1: Question answering over knowledge bases challenges and solutions (papers numbered in Appendix A).

Challenges	Solutions
Answering complex questions	Hybrid systems (3, 5). Graph similarity (62). More SPARQL modifiers compliance (53). Query ranking (52, 57, 64). Question paraphrasing (1). Seq2Seq (66). Siamese CNNs (17). Simple query composition (32, 56). Subgraphs matching (10, 51, 61). Templates (46, 63). Unsupervised message passing (58).
Answering simple questions	BERT transformer (59). CNN (29). Formal logic (2, 13, 14). Seq2Seq (15, 16, 18, 20, 21, 27, 35, 54). Simple pipeline (9, 11, 12, 23, 37, 47). Templates (4, 6).
Entity Linking	BERT transformer (60). Distant supervision (7). Joint entity and relation linking (41).
KB incompleteness	Hybrid system (19, 38, 65).
Modular design, module reusability	Integration framework (22, 36). Modules collection (43). Optimal module selection (48).
Relation Linking	BERT transformer (60). Distant supervision (33, 34). Hierarchical RNN (30). Joint entity and relation linking (41). LSTM (40, 45). Siamese LSTM (49, 55).
System tunings	User interaction (31). User interface (39). Query builder module (42, 50).
Training data scarcity	Automatic labelling (25, 28). Distant supervision (24, 26, 44). Multi-column CNN (8).

Table 3.2: Remaining challenges, future work (papers numbered in Appendix A).

Challenges. Future work	Research Directions
Answering complex questions (8, 15, 10, 16, 19, 21, 25, 42, 45, 53, 54, 57, 59, 64)	Conversational agent (21). Data augmentation (57). More SPARQL modifiers compliance (53). More training data (8, 10). Reinforcement learning (64).
KB incompleteness (26, 35, 49, 55, 60)	Hybrid system (26). More external knowledge; more training data (35, 49, 55, 60).

3.4 Summary

This systematic study collected information on the methods and challenges of QA over KBs, a topic that has gained traction in the search engine industry. The analysis of 66 papers allowed the classification of KBQA systems according to their architectural styles. Twenty-five semantic parsing pipeline systems were reported, as well as 12 using subgraph matching and seven based on templates. Twenty-two systems performing information extraction were also presented. The challenges ahead were presented, and some directions for future research were identified. Two primary challenges remain that are particularly sensitive to the success of this technology. On the one hand, it is necessary to answer increasingly complex questions; on the other hand, it is necessary to deal with the natural incompleteness of KBs. This study concluded that hybrid systems and adopting advanced machine learning techniques promise significant advances in the field.

Chapter 4

SCALEUS-FD: A FAIR Data Tool

Semantic annotations in knowledge management empowered the scientific community with solutions that make the most of distributed and heterogeneous data. The subject-predicate-object representation, together with ontologies, enables the annotation of knowledge and the creation of semantic repositories that can be massive. Additionally, the Findable, Accessible, Interoperable, and Reusable (FAIR) principles established guidelines for data sharing, gaining traction in data stewardship. However, one must deliver solutions smoothly integrated into the FAIR Data ecosystem to explore their full potential.

This chapter introduces SCALEUS-FD¹, a FAIR Data extension of a legacy semantic web tool for data integration and semantic annotation and enrichment. SCALEUS-FD enables online FAIR-compliant exposure of data and metadata by creating endpoints for machine-to-machine interactions. Deployed instances are self-descriptive and can be catalogued and found using search engines. Concepts revolving around the FAIR initiative are presented, as well as the software architectural details and implementation. Finally, a set of metrics allows evaluation of the tool's FAIRness.

4.1 FAIR Data Principles

The FAIR Data principles proposed by Wilkinson et al. (2016) provide guidelines to ensure that humans and machines can discover and reuse data resources. Not constrained by implementation decisions, the idea is to be as broad as possible, summarising the experience and best practices of the multiple institutions and individuals involved in research data sharing (Mons et al., 2017). A persistent identifier must be assigned to data and metadata and must be ensured to be indexed or registered in a searchable resource. Relevant attributes meeting domain-relevant community standards must be

¹<https://github.com/bioinformatics-ua/scaleus-fair>

used. Data and metadata use a formal language for knowledge representation and use vocabularies that follow FAIR principles. Data and metadata can be retrieved using a standardised communications protocol allowing authentication and authorisation when required. Furthermore, metadata should remain accessible even if data is no longer available. As stated by Wilkinson et al. (2016), explicitly, the principles are:

To be **Findable**:

- F1.** (meta)data are assigned a globally unique and persistent identifier;
- F2.** data are described with rich metadata (defined by R1 below);
- F3.** metadata clearly and explicitly include the identifier of the data it describes;
- F4.** (meta)data are registered or indexed in a searchable resource.

To be **Accessible**:

- A1.** (meta)data are retrievable by their identifier using a standardized communications protocol;
 - A1.1.** the protocol is open, free, and universally implementable;
 - A1.2.** the protocol allows for an authentication and authorization procedure, where necessary;
- A2.** metadata are accessible, even when the data are no longer available.

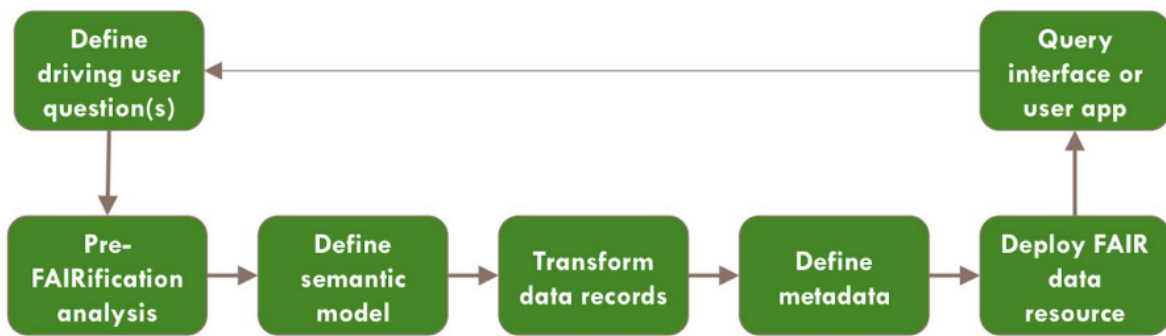
To be **Interoperable**:

- I1.** (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation;
- I2.** (meta)data use vocabularies that follow FAIR principles;
- I3.** (meta)data include qualified references to other (meta)data.

To be **Reusable**:

- R1.** meta(data) are richly described with a plurality of accurate and relevant attributes;
 - R1.1.** (meta)data are released with a clear and accessible data usage license;
 - R1.2.** (meta)data are associated with detailed provenance;
 - R1.3.** (meta)data meet domain-relevant community standards.

FAIRification work is not trivial and usually demands close collaboration between IT and domain experts. Although FAIR is not equal to RDF, LD, or SW, these technologies are a mature option for creating FAIR data (Mons et al., 2017; Wilkinson et al., 2017). Using the workflow shown in Figure 4.1, proposed by Jacobsen et al. (2018), helps convert data into FAIR.



Source: Jacobsen et al. (2018)

Figure 4.1: FAIRification steps: driving question(s) definition, pre-FAIRification analysis, semantic model definition, data records transformation, metadata definition, deployment, and query interface provision.

Several steps can be considered, starting with formulating domain questions and a pre-FAIRification analysis to focus and confront the original data with the desired outputs. The next step is to look closer at the data elements and define a semantic model capturing the domain experts' most relevant concepts and relations. One can reuse, adapt, combine, and augment existing models. The original data records are transformed to obtain a FAIR-compliant machine-readable representation by applying the developed ontological model. Then, the metadata about the data usage license and provenance in a format meaningful to computers is defined. Finally, after deploying the FAIR data resource, a query interface or user app is made available to end-users.

Some examples describing efforts to FAIRify life science data repositories can be reported. For instance, Rodríguez-Iglesias et al. (2016) present the FAIRification of a portion of the Pathogen-Host Interaction Database (PHI-base). Schaaf et al. (2018) report the extension of the Open Source Registry for Rare Diseases (OSSE) architecture to comply with FAIR principles, consisting of integrating a new component to expose metadata. Outside life sciences, can be highlighted the experiments of Garcia-Silva et al. (2019) around several Earth science disciplines.

4.2 Requirements and Building Blocks

In this section, the requirements are stated, and the building blocks of the solution are presented.

4.2.1 System Requirements

The ideas presented in the previous section lead to the following requirements:

Functional Requirements

- **It must be possible to store and describe multiple datasets** - The ability to store different datasets increases flexibility considering multiple domains or particular views of some specific domain.
- **Authorisation** - The tool enforces the authorisation levels defined by the dataset owners when using the solution. All users can access the metadata. Only authorised users can create or modify the metadata.
- **It must allow data queries** - Compliance with a widely used standard query language is a must-have.

Nonfunctional Requirements

- **It should be a standalone application** - Typical users are not IT personnel, and this underlines the need for the tool to be as simple to use as possible. The user's ability to start work immediately, skipping confusing configuration settings, is of paramount importance. If needed, the configuration process must be straightforward and well-documented.
- **It must be self-describing** - The solution must be by itself a FAIR object in the FAIR ecosystem. At the software level, metadata describing the deployed instance must be rich and preferably standard to allow the running solution to be registered and integrated into larger data interoperability systems.
- **It should make the data FAIRer** - Data resulting from tool processing should be as FAIR as possible.
- **It must expose its services over the web** - The tool must offer access points for other software agents to interact in a networked environment, fulfilling findability and accessibility criteria. Software agents access the data using a standardised communications protocol, allowing authentication and authorisation if required.
- **User-friendly interfaces** - Users are provided with a dashboard to see the stored datasets at a glance.

4.2.2 SCALEUS

SCALEUS² is a semantic web tool for data integration, validated in the scope of rare diseases (Sernadela et al., 2017b). The solution enables migration to a semantic format

²<https://github.com/bioinformatics-ua/scaleus>

without forcing using a predefined data integration ontology. This degree of freedom gives users greater flexibility in managing their data models. RDF resource loading is also available. Data is manageable as a collection because the tool supports the creation of multiple datasets. Another significant advantage is that people can quickly deploy and start using the single package software distribution, with no wasting time configuring.

The system enables users to perform a text search or SPARQL queries with inference rules to retrieve the stored information. Additionally, a simplified REST API allows several operations with different degrees of granularity, ranging from the dataset level to the level of the single triples. It is also possible to add, obtain, and remove namespaces. More importantly, a SPARQL endpoint is available for receiving and processing SPARQL queries over the web. In summary, the list of essential features is:

- 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.
- Web services API.

4.2.3 Data and Metadata FAIRness

Metadata establishes how data can be accessed and reused. A FAIR Data Point (FDP), as was proposed by Bonino da Silva Santos et al. (2016), provides a mechanism for users to discover properties (metadata) of datasets. The FDP is a central piece of the FAIR Data infrastructure, allowing the exposure of metadata in intermediate granularity between fully centralised descriptions of a super-collection of datasets or a fully distributed scenario where the metadata of each dataset is published individually. Metadata clusters with pointers to several datasets streamline indexation, registration, and search.

Indexing the solution's entry points in a search engine is paramount for data to become findable. It is essential to identify which search engines are most suitable for the

purposes that are intended. Implemented to scale to all metadata published on the web, the Google Dataset Search³ is a novel way to search for data collections automatically indexed by Google crawlers (Brickley et al., 2019). So, the solution must expose, using RDFa, Microdata, or JSON-LD, a description of the entry points for the datasets using the Dataset or the DataCatalog types from the Schema.org⁴ vocabulary. Another possibility is to use the Dataset concept from the W3C Data Catalog Vocabulary (DCAT) (Maali and Erickson, 2014). Adding simple markup describing datasets to web pages removes the need to build or directly feed a specific search engine and allows data exposure to a broad audience.

4.3 SCALEUS-FD

SCALEUS-FD is a semantic data publishing solution that follows the FAIR principles, as explained in this section.

4.3.1 Architecture of SCALEUS-FD

SCALEUS-FD is built on top of the legacy tool. As presented in Figure 4.2, the left branch of the architecture includes the SCALEUS components dealing with the process of semantic data conversion, and the right side shows the new elements of SCALEUS-FD, which allow the creation and management of metadata.

The components of the solution fall into three main layers: knowledge base, abstraction, and services. At the knowledge base layer, the databases store the datasets converted into semantic graphs by the users. At this same level, another triplestore stores the metadata as RDF triples, ensuring logical and physical separation between different types of data. The transaction database component (TDB) ensures that data are protected against corruption when dealing with creating, reading, updating, and deleting (CRUD) operations. The abstraction layer deals with managing semantic datasets at a higher level, comprising the methods for creating and manipulating the data and metadata. Finally, at the service layer, the tool exposes its functionalities through an API for machine-to-machine (M2M) interaction and a graphical user interface (GUI) for human clients.

The Data Handler provides the operations for converting the user's data into the semantic format. Metadata describing each of the created datasets must be entered or automatically generated and saved in the system. The ownership, license, and explicit description of the access points allow data navigation, fulfilling FAIR principles by

³<https://toolbox.google.com/datasetsearch>

⁴<https://schema.org>

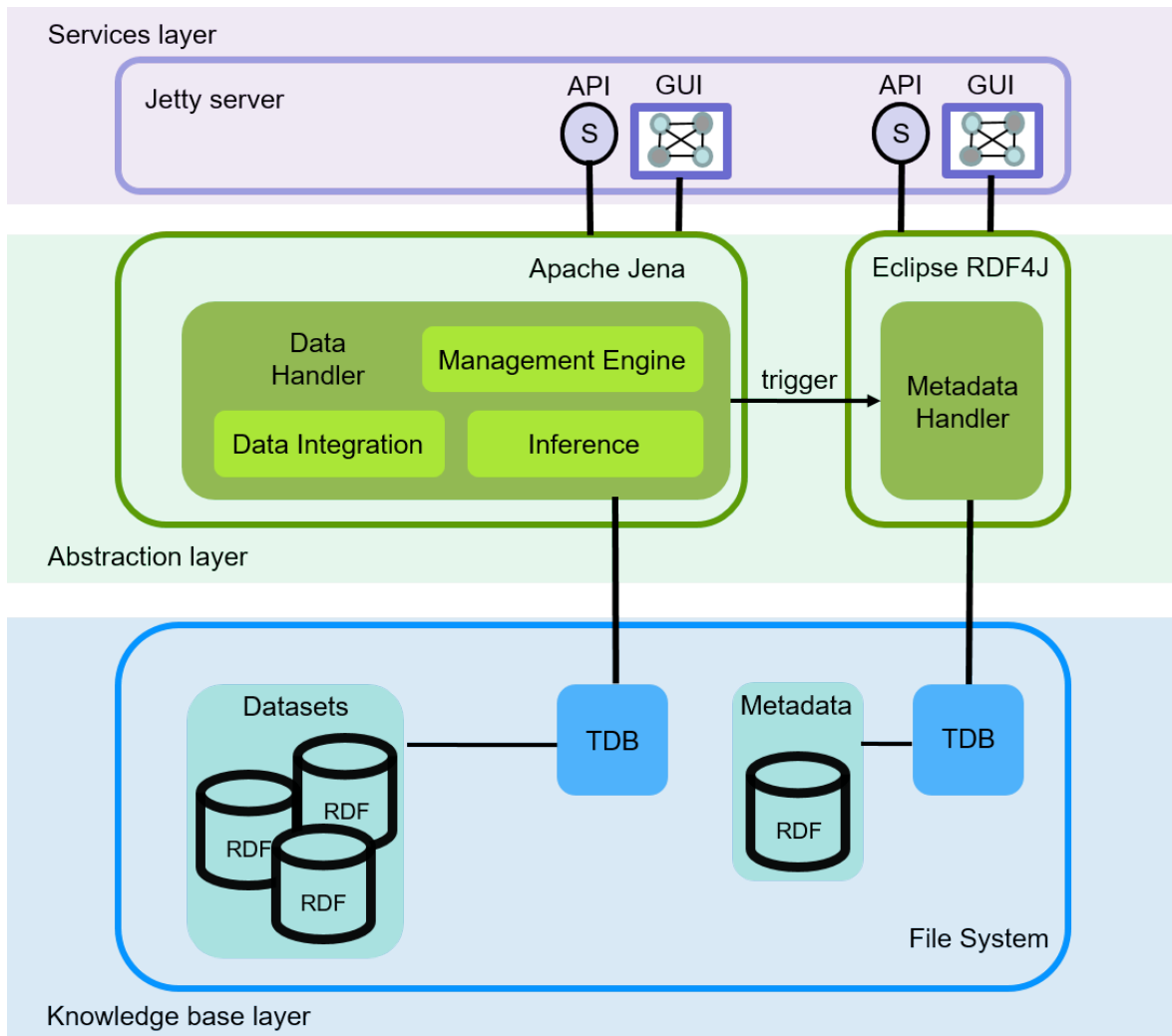


Figure 4.2: SCALEUS-FD architecture and implementation technologies. At the file system level are the triplestores for the converted data and the metadata. At the abstraction layer, Apache Jena and Eclipse RDF4J were used to implement the modules for dealing with the semantic data, comprising data integration, inference, and the management engine. Finally, a Jetty server allows the building of the services layer.

making reuse possible. Management of these metadata in semantic format is through the Metadata Handler component, which connects the TDB dealing with the metadata repository. The Data Handler module can directly trigger this module, although the metadata is also available via the services API and the GUI.

4.3.2 Metadata Hierarchy

Users can navigate between levels after clicking on any entry point exposed by a search engine, exploring the hierarchical metadata organisation. Figure 4.3 shows the metadata classes used to describe the tool, catalogues, datasets, and distributions. For

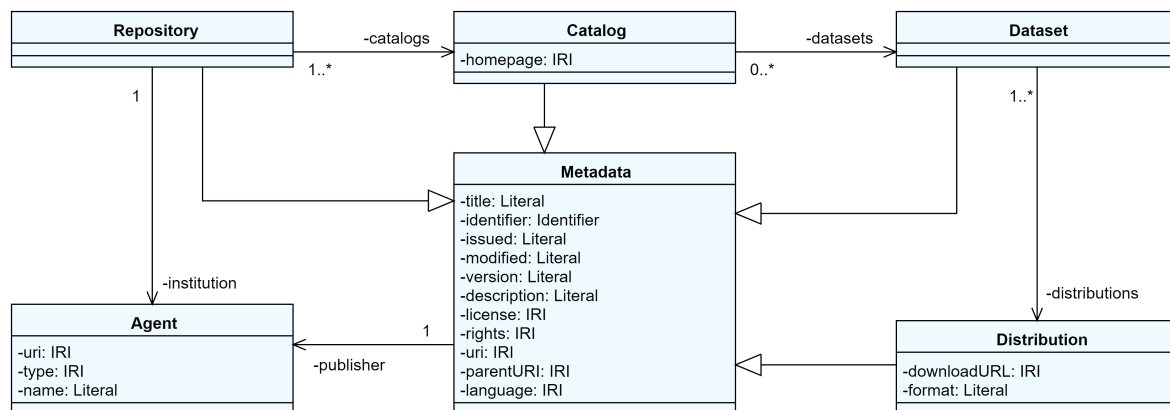


Figure 4.3: SCALEUS-FD metadata.

the profile, four levels of metadata are being considered, using the RE3Data Schema (Rücknagel et al., 2015) and the DCAT specification as a basis.

The first level of metadata describes the tool itself as a repository. By default, on the first run of the application, only one catalogue is created, but more can be added using the REST API. Users can change the default values for the first two layers using property configurations. In the third layer of metadata, a form is used to set the information about each added dataset. Finally, the distribution level is automatically created containing a data access URI.

4.3.3 Implementation

JavaScript libraries like AngularJS⁵ and CSS frameworks like Bootstrap⁶ were used to build a responsive web app. AngularJS is a JavaScript-based open-source front-end web framework for developing single-page applications. The backend modules use a standalone Eclipse Jetty⁷ web server and javax.servlet container. Jersey⁸ was used to implement RESTful web services complying with JAX-RS API⁹.

The Apache Jena¹⁰ solution was used to write and extract data from RDF graphs. Apache Jena is an open-source SW framework for Java. It provides an API to extract data from and write to RDF graphs. The fairmetadata4j¹¹ library was used to support the creation, storage, and provision of FAIR metadata. For metadata management,

⁵<https://angularjs.org/>

⁶<https://getbootstrap.com/>

⁷<https://www.eclipse.org/jetty/>

⁸<https://jersey.github.io/>

⁹<https://jcp.org/en/jsr/detail?id=370>

¹⁰<https://jena.apache.org/>

¹¹<https://github.com/FAIRDataTeam/fairmetadata4j>

the FAIRDataPoint¹² and Eclipse RDF4J¹³ were used.

4.3.4 Web Services API

A set of RESTful web services provides data and metadata management endpoints for external software applications, enabling M2M interaction. For instance, a dataset can be created or removed, or a list of all existing datasets can be retrieved. The same type of operation is available for namespaces management and at the level of triples. Creating, obtaining, or changing the tool’s metadata is also possible by evoking services (for more details, see the README file that comes with the source code on GitHub). More importantly, a generic SPARQL endpoint allows querying data and metadata unleashing the power of the SW approach.

4.4 Validation

This section presents a formal evaluation of the tool and its instantiation, considering a use case.

4.4.1 FAIR Maturity Assessment

A design framework and exemplar metrics to evaluate the FAIRness of any digital object were proposed by Wilkinson et al. (2018), considering the multidimensionality of the FAIR principles. Not only should data be evaluated but also any tool of the ecosystem must be FAIR compliant. Another important aspect is that this general framework of FAIR maturity indicators can be complemented with more specific assessment criteria to address the particular needs of particular communities. Next, the FAIRness assessment of the tool using the mentioned maturity metrics is presented.

- F1.** The rules of the “Persistent Domains” document presented as a design issue at <https://www.w3.org/DesignIssues/PersistentDomains.html> can be followed. HTTP URIs can be used to identify digital resources.
- F2.** With DCAT, data can be described considering different layers of machine-readable metadata.
- F3.** SCALEUS-FD’s metadata model allows setting a globally unique and persistent identifier for each digital resource.

¹²<https://github.com/FAIRDataTeam/FAIRDataPoint>

¹³<https://rdf4j.eclipse.org/>

- F4.** RDFa is used to embed the `dc:Dataset` class instances within the web documents generated by the app, allowing automatic indexing by the Google Dataset Search engine.
- A1.** See the assessment of A1.1-2.
 - A1.1.** Data and metadata are retrievable using HTTP, which is a free and open-source protocol.
 - A1.2.** The application provides basic access authorization to perform REST calls that create, update, or delete data and metadata (POST, PUT, and DELETE operations).
- A2.** After removing any dataset, metadata continues available.
- I1.** The RDF data model and the OWL formal language for knowledge representation were used.
- I2.** Datasets can be described using existing, well-known ontologies such as the HPO or GO. For the metadata, the DCAT vocabulary was used.
- I3.** Following the SW principles, ontologies that include semantically rich relationships were used.
- R1.** See the assessment of R1.1-3.
 - R1.1.** Accessible usage license: the “license” property of the `dc:Distribution` class is used to specify the license document by which the distribution is made available.
 - R1.2.** The `dc:Catalog` class keeps the information about data provenance.
 - R1.3.** SW standards for data and metadata are used.

4.4.2 Huntington’s Disease Use Case

The tool was used to increase the “FAIRification” of a registry with anonymised data from a cohort of patients with Huntington’s disease (HD), a fatal neurodegenerative disease affecting the brain. The source of information was a spreadsheet collecting genetic and phenotypic data from 151 patients. For the sake of security and privacy, this cohort’s data has been anonymised. Tabular data is a widespread format in the long tail of science and technology, and the small number of records is usual in the context of a rare disease, further underlining the importance of “FAIRifying” this data.

The data headers relate to enrolment (e.g., date of informed consent), demographics (e.g., gender), genetic testing results (e.g., CAG larger allele), medical history, comorbid conditions, and cognitive data columns related to the Problem Behaviours Assessment (PBA-s) items (McNally et al., 2015). Figure 3 shows the interface for loading the data to be converted to the semantic format.

The screenshot shows the SCALEUS Spreadsheet Integration interface. On the left is a sidebar with navigation options: Dashboard, SPARQL, Text Search, Namespaces, Triples, Data Preview, RDF Upload, and Spreadsheet. The main area is titled 'Spreadsheet Integration' and shows a file named 'Huntington Disease.xlsx' loaded. Below the file name is a table with columns: subject, created, site, subject_state, visit, svstdtc, visit_state, visi, and nr. The table contains 11 rows of data, each with a unique identifier, a date, a site name, a state, a visit date, a visit state, a visit count, and a number of records (nr).

	subject	created	site	subject_state	visit	svstdtc	visit_state	visi	nr
1	543-931-234	7/17/13	test-site	completed	Baseline	7/31/13	done	4579	1
2	953-252-345	10/8/04	test-site	violator	Baseline	5/31/12	done	3527	1
3	142-543-456	1/24/08	test-site	completed	Baseline	10/24/12	done	3830	1
4	343-274-567	2/23/06	test-site	completed	Baseline	11/21/11	done	3133	1
5	436-705-678	6/15/05	test-site	completed	Baseline	1/25/12	done	3258	1
6	922-026-789	9/14/04	test-site	completed	Baseline	8/14/12	done	3665	1
7	235-097-890	2/23/07	test-site	completed	Baseline	8/25/11	done	2960	1
8	084-618-901	3/6/08	test-site	completed	Baseline	7/4/12	done	3590	1
9	336-081-234	12/9/04	test-site	completed	Baseline	8/24/11	done	2958	1
10	822-302-345	10/19/12	test-site	completed	Baseline	10/19/12	done	3816	1
11	903-123-456	2/2/05	test-site	completed	Baseline	10/9/12	done	3811	1

Figure 4.4: Spreadsheet integration interface.

After loading the data, the columns to be transformed into the semantic format are selected. For each column, one must associate the semantic entity and namespace according to the selected ontologies. Concepts from the Dublin Core Metadata Initiative¹⁴, FOAF Vocabulary Specification¹⁵, and the Human Phenotype Ontology¹⁶ were used. Table 4.1 shows the performed mapping.

Table 4.1: 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/

For instance, one can map the “subject” column to the term <http://purl.org/dc/terms/identifier/> from the Dublin Core Metadata Initiative, and the “gender” column to the property <http://xmlns.com/foaf/0.1/gender/> from the FOAF Vocabulary Specification. Other ontologies can be used, as the Human Phenotype Ontology (<https://hpo.jax.org/app/>) to map columns like “depression” (HP:0000716), “irri-

¹⁴<https://dublincore.org/>

¹⁵<http://www.foaf-project.org/>

¹⁶<https://hpo.jax.org/app/>

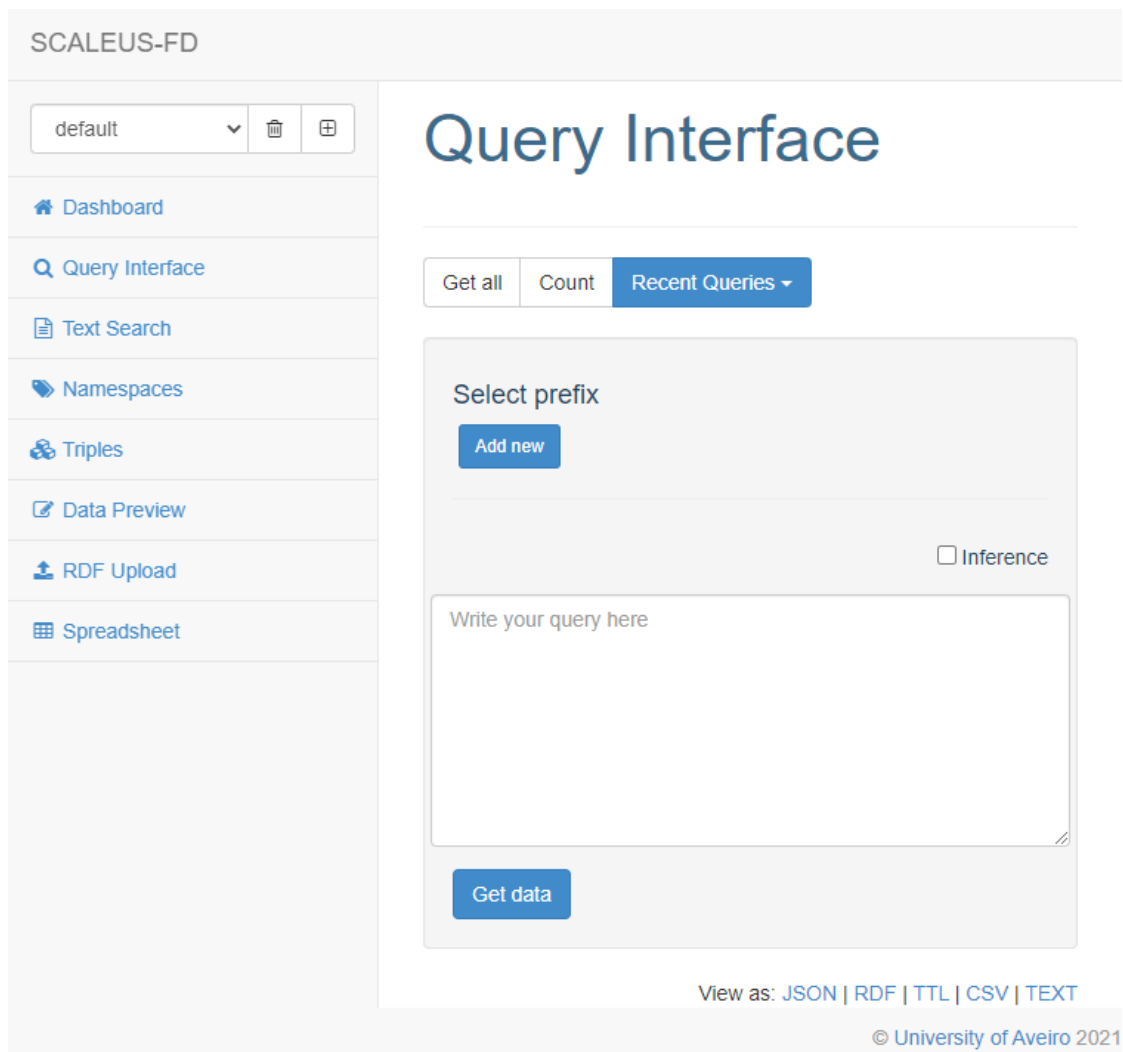


Figure 4.5: QA interface.

tability” (HP:0000737), “psychosis” (HP:0000709), and “apathy” (HP:0000741). The conversion process concludes by creating the triples that are loaded into the preselected dataset. With the data transformed and adequately loaded, one can ask questions using a graphical interface (see Figure 4.5).

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

4.5 Summary

SCALEUS-FD is a tool created to ease the burden of publishing FAIR-compliant data and metadata to facilitate interoperability and reuse. The solution uses the SW and LD principles, and its “FAIRness” has been assessed against a set of maturity metrics.

The solution has been validated in the field of rare diseases, proving to be a valuable aid for people looking for data sharing.

Chapter 5

Querying Semantic Data

The secondary use of health data is a valuable source of knowledge that drives observational studies, leading to important discoveries in the medical and biomedical sciences. For example, observational studies suit pharmacological surveillance, public health monitoring, expanding knowledge about endemics and epidemics, and development of new treatments (Schneeweiss and Avorn, 2005). The fundamental guiding principle for conducting a successful observational study is carefully formulating the research question and the data search approach. However, finding and integrating suitable datasets to support multicentre studies is challenging, time-consuming, and not infrequently impossible without a deep understanding of each dataset (Nan et al., 2022).

This chapter presents a strategy for retrieving semantically annotated biomedical datasets, using an interface built by applying a methodology to transform natural language questions into formal language queries (Figure 5.1). Using natural language interfaces to issue complex questions without directly manipulating a logical query language enhances the advantages of creating and using biomedical semantic data.

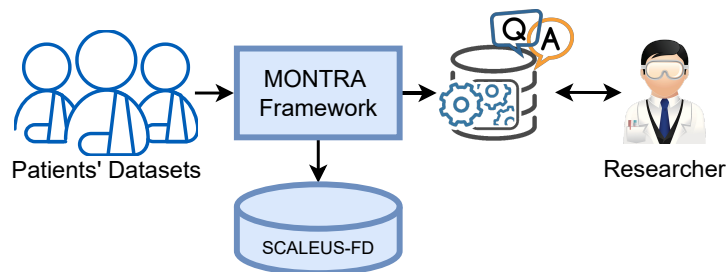


Figure 5.1: Overview of question answering over semantic biomedical data. The approach allows the publishing of patient datasets metadata in a biomedical database catalogue built using the MONTRA (Silva et al., 2018) framework. SCALEUS-FD operates as a FAIR repository of ontologies. Researchers can consult the data using a built-in question-answering module described in this chapter.

The methodology was validated considering a use case based on Alzheimer’s disease datasets published on a European platform for sharing and reusing biomedical data. Data were converted to semantic information format using biomedical ontologies in everyday use in the biomedical community and published as a FAIR endpoint. Three natural language question types for the biomedical semantic data were considered: single-concept, exclusion criteria, and multiple-concept questions. Finally, the performance and limitations of the developed question-answering module were analysed. The source code is publicly available at <https://bioinformatics-ua.github.io/BioKBQA/>.

In a nutshell, a strategy for using information extracted from biomedical data and transformed into a semantic format using open biomedical ontologies was proposed. The method uses natural language to formulate questions to be answered by this semantic data without directly using formal query languages.

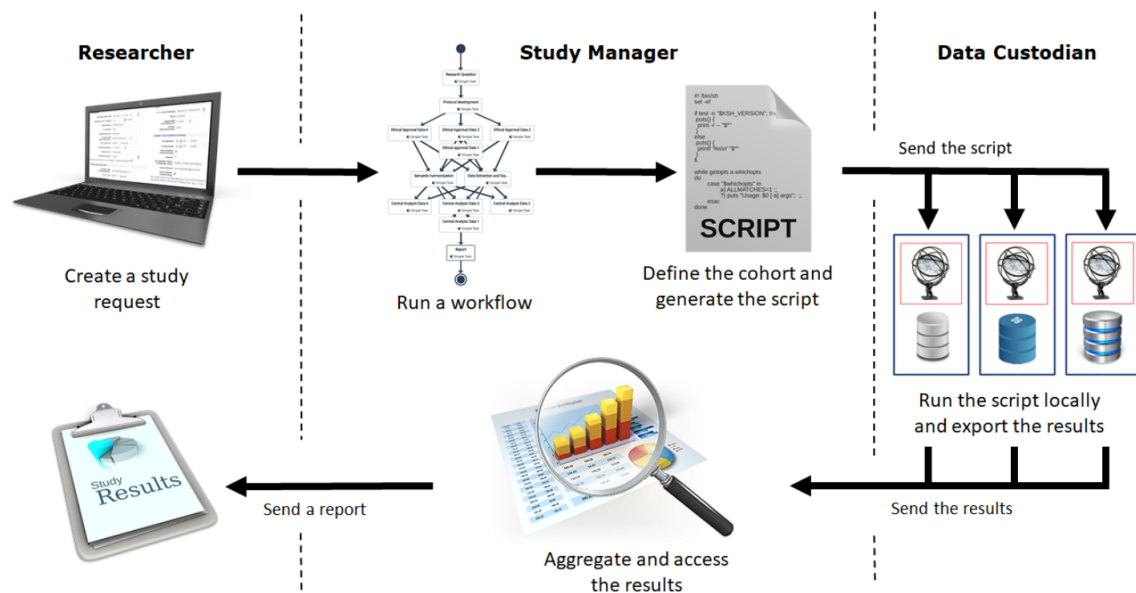
5.1 Contextualisation

The digitisation of medical information resulted in massive 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 (Cheng and Phillips, 2014). Over time, researchers worldwide have created repositories of biomedical data in various formats, from specialised databases to simple tabular data (Kolker et al., 2012). 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, busting new fields of study and significantly increasing the impact and recognition of research outputs (Wallis et al., 2013).

Different strategies to solve data sharing and interoperability problems can be pointed out. 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)¹ initiative (Hripcsak et al., 2015). 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 for extracting and transforming the original data into the new format to be loaded into a database or made available as tabular data. Natu-

¹<https://www.ohdsi.org/>

rally, 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 constraints (Francis and Francis, 2017). A strategy for publishing these databases' existence is based on characterising each dataset, using data aggregation and meta-data. 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. Then they can access the data using data access pipelines, such as Fajarda et al.'s (2018) pipeline depicted in Figure 5.2.



Source: Fajarda et al. (2018)

Figure 5.2: Query process workflow of common data model-based databases. A researcher creates a question to be processed by a study manager who scripts the necessary SQL queries using a work management system, such as TASKA (<https://bioinformatics.ua.pt/taska>) (Almeida et al., 2018). The data custodians run the script and forward anonymised results to the study manager, who compiles, aggregates, and sends them to the researcher, closing the loop.

An evolution of the previous solution, shown in Figure 5.3, uses semantic technologies to harmonise the different databases using semantic adapters (Almeida et al., 2019). However, users need to create SPARQL queries which are not easy to do by standard users.

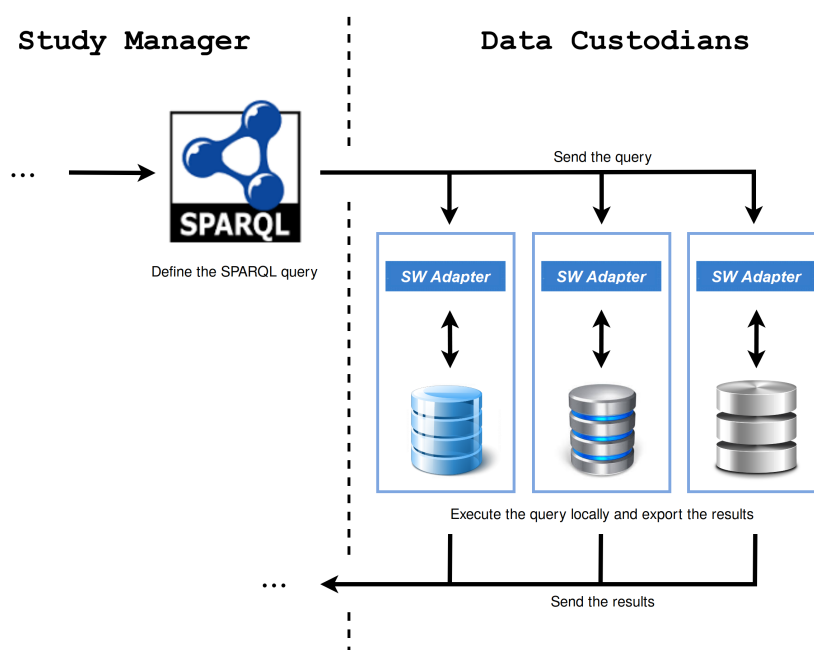


Figure 5.3: Query process workflow of semantically annotated databases. In this scenario, the study manager scripts SPARQL queries. A set of adapters interfaces with the databases using a pre-agreed ontology that assists in data retrieval.

Correctly selecting the study design and databases is essential to ensure the study’s feasibility. 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 (Höffner et al., 2017). Contrarily, the objective is simple data access without losing power in question formulation. A first approach to solving this problem can be using 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 (Ferré, 2017).

Question answering over knowledge bases allows asking natural language questions to obtain concise answers from semantic databases, freeing users from knowledge of the data schema and formal languages (Pereira et al., 2022). These systems rely heavily on advanced natural language processing techniques and are constantly evolving to accommodate increasingly complex natural language queries, as was surveyed in Chapter 3. However, their use for biomedical semantic data remains challenging because of lexical ambiguity, question abstraction issues, and query generation problems (Hamon et al., 2017).

5.2 Background

Semantic technologies allow researchers to share their data in a distributed and interoperable way. In this context, it is essential to know how to 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.

5.2.1 Discovery of Biomedical Databases

Searching for datasets raises different challenges from those faced with current web searches (Kern and Mathiak, 2015). 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 (Kacprzak et al., 2017). Other difficulties arise from 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 (Goel et al., 2010). 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.org² (Brickley et al., 2019).

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 Findable, Accessible, Interoperable and Reusable guidelines. The FAIR data principles intend to ensure that humans and machines can discover and reuse data resources (Wilkinson et al., 2016). 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 (Mons et al., 2017). Persistent identifiers must be assigned to data and metadata and guarantee registration in a searchable resource. One 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. The retrieval of data and metadata must be done 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.

Some examples of biomedical data discovery platforms can be pointed out. The application of semantic technologies is at the base of several platforms. BioSharing covers life science topics related to standards, databases, and policies (McQuilton et al., 2016).

²<https://schema.org/>

Also, YummyData (Yamamoto et al., 2018) is based on Linked Data to promote the discovery of biomedical databases and the Open PHACTS Discovery Platform (Groth et al., 2014) 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 (Sansone et al., 2017). The EHR4CR platform integrates clinical data from several hospitals and pharmaceutical companies in seven European countries (De Moor et al., 2015). 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 (Oliveira et al., 2019).

5.2.2 Managing Biomedical Data with Semantic Web Technologies

Semantically organised data present a logical structure that facilitates inferring new knowledge, and can be used directly to answer questions (Fan et al., 2012). Therefore, it is convenient to store the knowledge extracted from structured or unstructured data in a Knowledge Base (KB). The data of a KB can be considered to be organised as an edge-labelled multidigraph (Paulheim, 2017). 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) (Schreiber and Raimond, 2014). Based on this simple data model, one can build more complex models by semantic extension.

Standard vocabularies and ontologies allow modelling shared conceptualisations of knowledge domains by establishing classes, properties, individuals, and data values (Borst, 1997). Some notable contributions regarding life sciences can be pointed out. The Human Phenotype Ontology (HPO) vocabulary describes human diseases' phenotypic abnormalities (Köhler et al., 2016). The Orphanet Rare Disease Ontology (ORDO) is a resource for annotating rare disease data that provides relationships between relevant traits, namely diseases and genes (Weinreich et al., 2008). Gene Ontology (GO) describes genes considering molecular functions, cellular components, and biological processes (Gene Ontology Consortium, 2016). The ELIXIR³ (European Life Sciences Infrastructure for Biological Information) initiative also offers an ontology repository platform (Drysdale et al., 2020). Many more biomedical ontologies and ter-

³<https://elixir-europe.org/>

minologies are available on the BioPortal repository (Whetzel et al., 2011), 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 on five areas: 1) register and benchmarking 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-Connect⁴ initiative created an infrastructure for rare disease research to improve the analysis and sharing of genomic data, patient registries, and virtual biobanks (Thompson et al., 2014). The Biodiversity Community Integrated Knowledge Library (BiCIKL)⁵ 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 (Penev et al., 2021).

5.3 Materials

The proposal aims to add new functionality to search semantic data in natural language. This work seeks to improve a legacy biomedical data catalogue solution and uses a previously developed ontology repository.

5.3.1 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 Catalogue⁶ (Oliveira et al., 2019), a FAIR platform where data custodians can publish metadata about their biomedical databases with different levels of granularity (Trifan and Oliveira, 2018). This catalogue used the MONTRA⁷ framework (Silva et al., 2018) to allow the publishing and discovering of data.

⁴<https://rd-connect.eu/>

⁵<https://bicikl-project.eu/>

⁶<https://emif-catalogue.eu/>

⁷<https://github.com/bioinformatics-ua/montra>

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, a new metadata search module can be added, improving the base search capabilities. The solution also incorporates a REST API that allows interactions with third-party software applications.

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. The query in its simplest version can be built 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. One also has a form with all possible fields, with which can be built 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.

The use of questions in natural language is an asset for users because it allows the construction of complex queries without prior knowledge of the data structure. The proposal that enables a natural language interface was developed as a MONTRA plugin. In addition to the classic form-based search methods, now there is an easier and more intuitive way of searching for databases described in a catalogue.

5.3.2 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 (Pereira et al., 2020).

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, it is needed to ensure they can be accessed using open com-

munication 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 as an ontology repository ensures all these desirable FAIR characteristics, as assessed by Pereira et al. (2020) using the maturity metrics proposed by Wilkinson et al. (2018).

When using metadata to describe catalogues, it is established how 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)⁸. SCALEUS-FD uses RDFa to enable web crawlers to index DCAT annotations automatically.

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

5.4 Methods

A semantic data questioning system using natural language and its integration in a biomedical database catalogue solution (Figure 5.4) is proposed. The solution includes several phases, starting with the creation of lexicons of entities and relationships. These lexicons are then used in the subsequent two 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 the pipeline and are further detailed in Section 5.4.2.

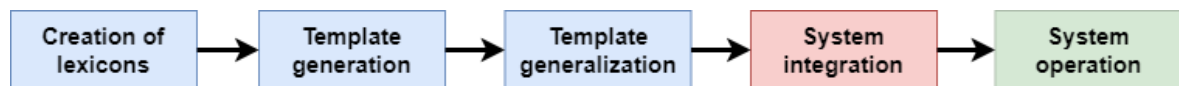


Figure 5.4: General overview of the QA approach.

⁸<https://www.w3.org/TR/vocab-dcat-2/>

5.4.1 Natural Language Queries over Knowledge Bases

Considering the pairwise disjoint sets \mathcal{I} of IRIs, \mathcal{B} of blank nodes, and \mathcal{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 $V_1 = \mathcal{I} \cup \mathcal{B}$, $V_2 = \mathcal{I} \cup \mathcal{B} \cup \mathcal{L}$, and a labelled arc set $E^* = \{(v_1, l, v_2) : v_1 \in V_1, v_2 \in V_2, l \in L\}$, l being an element of the label set $L = \mathcal{I} \cup \mathcal{B} \cup \mathcal{L}$. 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 $V = C \cup In \cup \mathcal{L}$, where C is a set of classes, In is a set of class instances, and \mathcal{L} is a set of literals. A multitude of predicates can exist connecting two nodes. Each pair of nodes plus the connecting predicate is called a fact. A path is a sequence $(v_0, a_1, v_1, \dots, a_n, v_n)$, $n > 0$, alternating nodes $(v_i, i = 0, \dots, n)$ and labelled arcs $(a_j, j = 1, \dots, n)$. 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 . Nodes representing n-ary relations can also be considered to accommodate more complex cases, coded by creating an individual representing the relation instance itself or using an RDF vocabulary for lists.

Creation of Lexicons

The first step is the construction of 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, a lexicon Lex_e was created 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 (Daiber et al., 2013).

To build Lex_e , each (e_1, r, e_2) triple is used to 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, being added to the lexicon the mappings $\{syntactic_unit_1 \rightarrow e_2, syntactic_unit_2 \rightarrow e_1\}$. It is followed a similar principle to construct the predicate lexicon. For this set, considering each (e_1, r, e_2) triple, the pattern $\langle e_1 syntactic_unit e_2 \rangle$ is identified and the mapping $\{syntactic_unit \rightarrow r\}$ is added to Lex_r . Note that more patterns can be added later to increase the system's sensitivity.

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 to answer questions at testing

created automatically and use the full KB type system as potential mapping targets. Starting with \hat{q} generated thus far, the answer variable node in \hat{q} is connected to one type constraint for each $c \in C$ such that the variable originates from the answer entity $a \in A_u$ and $(a, type, c) \in KB$ (Figure 5.7).

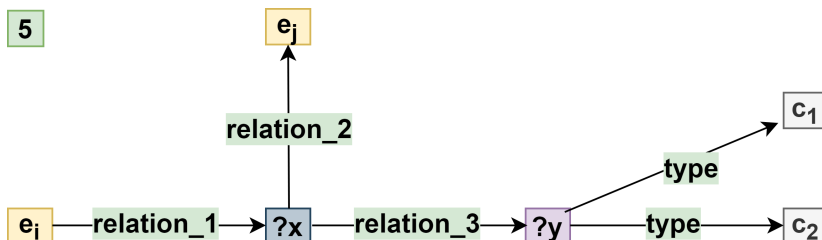


Figure 5.7: Looking at the answer a (currently variable $?y$), it is determined to which classes it belongs as an instance. Two classes are shown in the figure, but of course, the number of classes can be different.

With (u, \hat{q}) pairs at hand, the constituents of u and \hat{q} are aligned. The alignment gives the chunking of u into phrases that map to semantic items in \hat{q} . Alignment is driven by lexicons Lex_e and Lex_r (see Figure 5.8), but faces inherent ambiguity, either from truly ambiguous phrases or from noise in the automatically constructed lexicons. The resolution of this ambiguity is modelled as constrained optimisation and uses Integer Linear Programming (ILP) to address it.

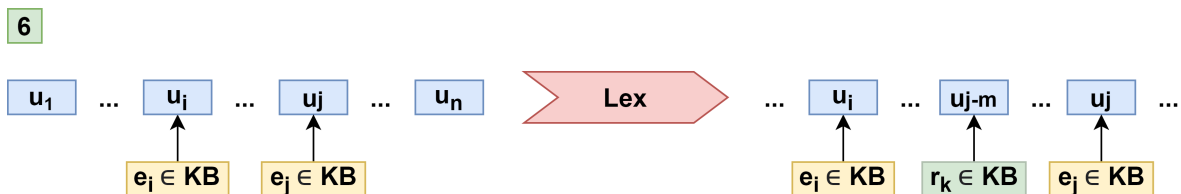


Figure 5.8: This step applies the entity and relationship lexicons to find relationships between entities and, possibly, some new entities.

A bipartite graph with Ph , the set of all phrases from u , is built on one side and $S_{\hat{q}}$, the set of semantic items in \hat{q} on the other. $Ph = ph_1, ph_2, \dots$ is generated by taking all subsequences of tokens in u . An edge is added between each $ph_i \in Ph$ and $s_j \in S_{\hat{q}}$ 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 one wants to pick at most one, this set of types is $S(v) = c_1, c_2, \dots$ and the set of phrases that can map to types in $S(v)$ is $Ph(v)$. Finally, to solve the ILP problem, IBM ILOG CPLEX

Optimizer⁹ is used, but other solvers can be integrated programmatically because the system is solver-agnostic.

Template Generalisation and System Operation

The aligned utterance-query pairs obtained from the alignment process are generalised. On the utterance side, the utterance u is represented using its dependency parse tree and restricted to the smallest connected subgraph that contains the tokens of all phrases participating in m . To create a template from this subgraph, the nodes participating in m are converted into placeholders by removing their text and keeping the POS tags and semantic alignment annotations ($ent, type, pred$). Universal POS tags are used for stronger generalisation power. Compound nouns are replaced with a noun token that can be used to match compound nouns at testing time to ensure generalisation. At testing time, the 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). On the query side, the concrete labels of edges (predicates) and nodes (entities and types) participating in m are removed from the query, keeping the semantic alignment annotations. The number of utterance-query pairs which generate a template is used as a signal in query ranking.

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

5.4.2 System Integration

The proposed KBQA applied to bio-databases reuses two open-source tools, avoiding the development of new components with similar goals. Therefore, the MONTRA Framework was adopted to integrate the tool as a plugin and the SCALEUS-FD to serve as an ontology repository. Figure 5.9 represents an overview of the architecture of the 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 component uses the NLP processor to perform the

⁹<https://www.ibm.com/analytics/cplex-optimizer>

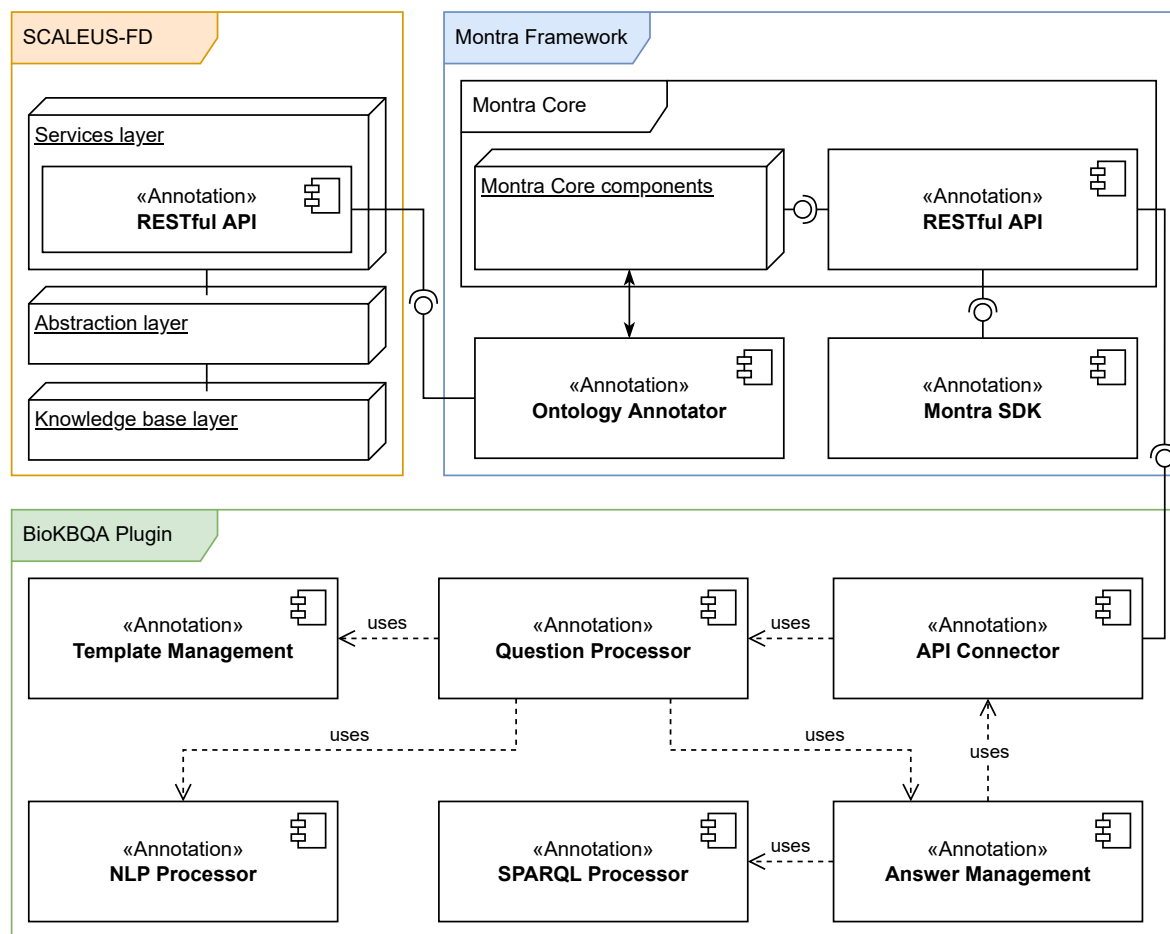


Figure 5.9: Component diagram showing the integration of SCALEUS-FD, MONTRA, and the BioKBQA plugin. The MONTRA block is a client of SCALEUS-FD that works as a repository of ontologies and of the BioKBQA plugin that allows querying ontologies using natural language.

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.

The 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.

Semantic Questioning

The question answering (QA) module added to SCALEUS-FD allows querying stored semantic data. On the one hand, it 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. The linguistic processing tools were integrated into the module to enable 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, one can use API calls that make it possible to retrieve information through software agents. Two endpoints were created 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. Figure 5.10 shows the offline and online phases of creating and using templates.

5.5 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.

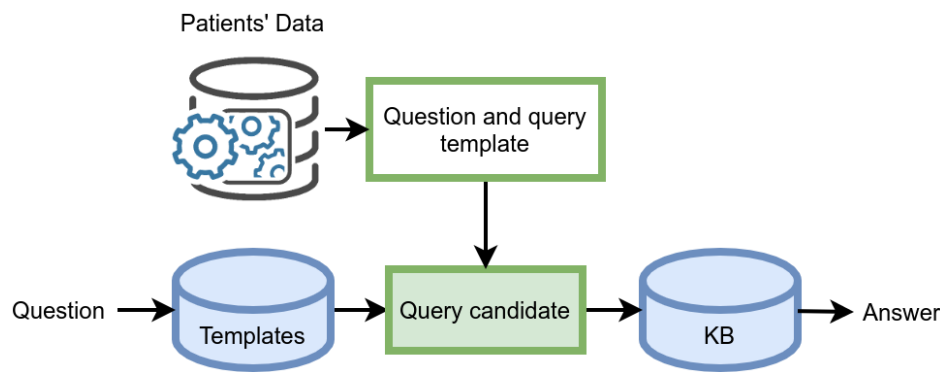


Figure 5.10: A high-level view of creating and using templates considering the offline and online phases. A corpus is processed in the offline phase to create pairs of natural language question templates and formal language query templates. Each question is disaggregated in the online stage, with the system running to determine the most suitable model.

The proposed methodology can be used by researchers to define simple cohorts over patient datasets, independently of the dataset used. This proposal’s main overhead is defining and mapping the fields and concepts in the database to 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 associates concepts in the data with their standard definition.

5.5.1 Use Case Overview

Simple models have been identified 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, three main categories of question-answering templates were defined in the 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 an ontology is defined to create metadata annotations about the datasets.

The European Medical Information Framework (EMIF)¹⁰ 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

¹⁰<http://www.emif.eu>

to accelerate the discovery and validation of new biomarkers to diagnose 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.

5.5.2 Ontology

This contribution follows from the work of the EMIF-AD project, where an ontology was constructed to annotate Alzheimer's disease data¹¹. 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. The ontology is based on the fields of this MONTRA-loaded questionnaire¹².

An ontology was built, reusing standard medical and biomedical ontologies and vocabularies, guided by the METHONTOLOGY methodological framework (Fernández et al., 1997). DCAT was used to annotate essential information about the repositories described on the platform. The DCMI Metadata Terms¹³ were used to annotate bibliographic resources. To report about clinical trials, the Ontology for Biomedical Investigations¹⁴ was used. To describe nuclear radiology entries, the RadLex radiology lexicon¹⁵ was used.

The name of the database was mapped to the DCAT property `http://purl.org/dc/terms/title`, and the `http://purl.org/dc/terms/accessRights` term provides access privileges and security status information. To insert a bibliographic reference, the term `http://purl.org/dc/terms/bibliographicCitation` was used. An exclusion criterion in a clinical trial was annotated with the term `http://purl.obolibrary.org/obo/OBI_0500028`. A magnetic resonance imaging (MRI) was annotated with the term `http://radlex.org/RID/RID10312`. Therefore, the ontology follows a hierarchical structure and is subdivided into the following 26 domains:

1. Database General Information: provides general information about the database, namely the name of the database, acronym, institutional data, and responsible people.

¹¹<https://bioportal.bioontology.org/ontologies/EMIF-AD/?p=summary>

¹²<https://github.com/bioinformatics-ua/BioKBQA/blob/master/resources>

¹³<https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>

¹⁴<http://purl.obolibrary.org/obo/obi.owl>

¹⁵<http://radlex.org/>

2. Key Publications: registration of publications containing relevant information about the study design, such as the number of participants and the techniques/instruments used. It is not necessary to be exhaustive in the bibliographic report.
3. Data Access: Indicates data sharing availability and whether additional informed consent or other procedures are required. Data sharing includes direct or mediated access after a research request submission and approval.
4. Inclusion/Exclusion Criteria: General categories of inclusion and exclusion, such as age group or pre-existing clinical conditions.
5. Number of Subjects: Estimated number of subjects available.
6. Clinical Information: Indicate what information is available in the data for analysis, such as educational status.
7. Dementia and Functional Rating Scales: The used dementia rating scales. Description of the measurement instruments, version (if applicable), frequency of application, and time interval between applications (monthly, annual, etc.).
8. Subjective Cognitive Impairment: The used subjective cognitive rating scales. Description of the measurement instruments, version (if applicable), frequency of application, and time interval between applications (monthly, annual, etc.).
9. Neuropsychiatric Scales: The used neuropsychiatric scales. Description of the measurement instruments, version (if applicable), frequency of application, and time interval between applications (monthly, annual, etc.).
10. Quality of Life: Quality of life assessment. Description of the measurement instruments, version (if applicable), frequency of application, and time interval between applications (monthly, annual, etc.).
11. Caregiver: Caregiver burden, impacts, and work status or productivity assessment. Description of the measurement instruments, version (if applicable), frequency of application, and time interval between applications (monthly, annual, etc.).
12. Health Resource Utilisation: Types and frequency of health care utilization, such as hospitalization. Frequency (an isolated case, recurrent cases), the time interval between occurrences (monthly, annual, etc.). Indication of the information collection method (self-report, electronic health records, another specific instrument, etc.).
13. Remote Monitoring Technologies: Usage of remote monitoring technologies, such as wearable devices, smartphone-based solutions, sensor-based technologies, and computer-based technologies.

14. Cognitive Screening Tests: The used cognitive screening tests. Description of the measurement instruments, version (if applicable), frequency of application, and time interval between applications (monthly, annual, etc.).
15. Neuropsychological Tests: The used neuropsychological tests. Description of the measurement instruments, version (if applicable), frequency of application, and time interval between applications (monthly, annual, etc.).
16. Physical Examination: Indication of what anthropomorphic measures have been utilized. Specification of how the assessment was made for a neurological examination or evaluation of extrapyramidal signs or symptoms. Indication of the data collection frequency and the time interval between measurements (monthly, annual, etc.).
17. Lifestyle Factors: Lifestyle factors measurements.
18. Blood Collection: Specification of blood collection, sera, or plasma and if DNA or RNA analyses were performed. Upload the procedure or protocol manual, or enter information about the details of the collection procedures.
19. CSF Collection: Specify if CSF was collected and analyzed. Upload the procedure or protocol manual, or enter information about the details of the collection procedures.
20. Urine Collection: Specify if urine was collected and analyzed. Upload the procedure or protocol manual, or enter information about the details of the collection procedures.
21. MRI: Specifies whether an MRI was performed only once or in multiple visits and the time interval between exams. Upload the procedure or protocol manual, or enter information about the details of MRI scanning procedures.
22. PET: Specifies whether a PET was performed only once or in multiple visits and the time interval between exams. Upload the procedure or protocol manual, or enter information about the details of PET scanning procedures.
23. CT Scans: Specifies whether a CT scan was performed only once or in multiple visits and the time interval between exams. Upload the procedure or protocol manual, or enter information about the details of CT scanning procedures.
24. SPECT Scans: Specifies whether a SPECT scan was performed only once or in multiple visits and the time interval between exams. Upload the procedure or protocol manual, or enter information about the details of SPECT scanning procedures.

25. Electrophysiology: Specifies if electrophysiology measures were performed only once or in multiple visits and the time interval between measurements. Upload the procedure or protocol manual, or enter information about the details of electrophysiology measures procedures.
26. Neuropathology: Specifies if neuropathology on autopsy was obtained. Upload the autopsy procedure or protocol manual, or enter information about the autopsy details.

FAIRness is guaranteed by the used tools. The EMIF Catalogue is a FAIR platform, as demonstrated by Trifan and Oliveira (2018). Likewise, the SCALEUS-FD used as an ontology repository is a FAIR tool, as assessed by Pereira et al. (2020) using the maturity metrics proposed by Wilkinson et al. (2018).

5.5.3 Use Case Examples

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, or 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 the methodology.

The starting point was “The Book Of Ohdsi”¹⁶, where a broad set of questions is formulated for the creation of cohorts from the consultation of database catalogues. Table 5.1 presents six examples of the 30 questions processed that fit into three defined categories for this research application: C1) Single concept question; C2) With exclusion criteria; C3) With Multiple concepts. The output can be provided as: O1) a summary of the data, which aggregates the information, usually a numeric count; or O2) patients’ data filtering and retrieval, providing a cohort of patients based on the question conditions.

5.5.4 Validation and Error Analysis

The proposed methodology seems promising in exploring semantic datasets, achieving an accuracy of 0.76, which results from the successful processing of 23 of the 30 questions considered. Studying the limitations, sometimes the error is because it is

¹⁶<https://ohdsi.github.io/TheBookOfOhdsi/DataAnalyticsUseCases.html#characterization>

Table 5.1: Examples of questions, divided into three main categories (C) with two outputs (O): C1) simple question; C2) question with exclusion criteria; C3) questions with more than one concept; O1) data aggregation; and O2) patients' data filtering and retrieval.

Output	Category	Question Example
	C1	How many patients performed the neuropsychological examination?
O1	C2	Amount of patients performed a PET exam but did not perform the auditory verbal learning test.
	C3	Number of patients that performed the animal fluency test in 1 and 2 minutes.
	C1	Which patients performed attention and MRI scan?
O2	C2	All the patients that performed the Boston naming test and WAIS?
	C3	Datasets with visuoconstruction and batteries tests.

impossible to map the relationship between two entities. This happens, for example, with the question “What test is recommended to detect Lewy Body Dementia?”. For this question, the disease is registered in the dataset, there are patients entered, and a mention of tests performed is found. The problem is that the “is recommended” relationship does not exist. So there is no correct triple that can be extracted from the database to get an answer.

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 forms into the semantic elements present in the database. For example, it was not possible to define a strategy capable of defining two sets of patients and comparing them. An example of a question related to the presented research use case would be “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.

5.6 Discussion

Creating metadata to describe biomedical databases allows researchers to find them in an integrated way. A wide range of mature tools can be used to create, maintain, and store ontologies to operationalise semantic operations. Tools like Protégé allow building the ontology to capture the knowledge domain. The conversion to semantic data can be performed using tools such as SCALEUS-FD, which can also be used 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. The 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 quickly.

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 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 discovering other databases and operating in a scenario of interoperability, as is possible with semantic technologies. So, once again, the proposal overcomes these difficulties because enables to search the metadata of database catalogues using a natural language interface for simplicity.

Figure 5.11 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, even if some duly documented recursion is admissible. Researching the data of interest is crucial for the study's success. A recommendation system that offers the datasets can be used at this stage (Almeida et al., 2020). However, this solution is not always flexible and effective as it depends on historical data. The 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.

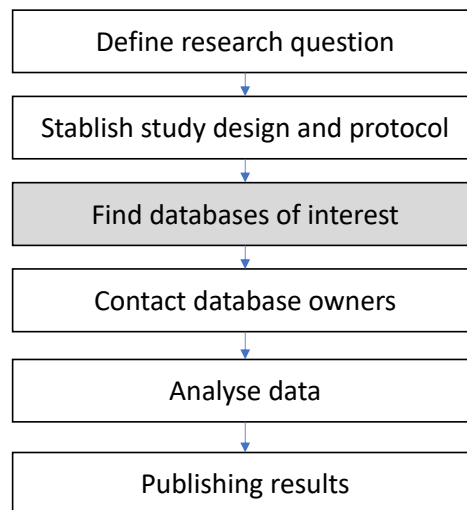


Figure 5.11: Typical observational study pipeline, from the research question definition phase to publication of the results.

Therefore, after specifying the study protocol, a researcher using the proposed system defines the question in natural language handled by the BioKBQA plugin, as shown in Figure 5.12. 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 form 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. 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.

5.6.1 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

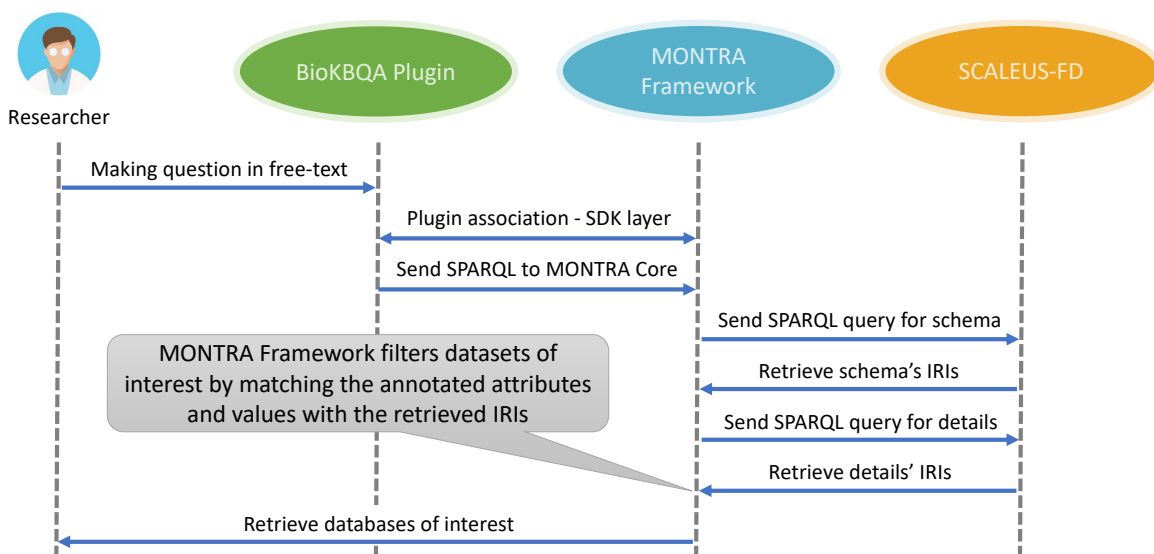


Figure 5.12: Interaction diagram showing the interactions between the different systems involved in answering a question asked by a researcher using natural language.

knowledge base. It is also interesting to increase the available data while simultaneously considering unstructured data, such as text. If this is the case, one is considering 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.

5.7 Summary

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

The proposed system extends the functionality of biomedical database catalogues to simplify searching for databases. So, in addition to the possibility of using forms to build queries, one can now use an interface that accepts questions in natural language. The method uses automatically constructed templates and is based on creating an ontology that is used to annotate the descriptors of the databases of interest. The

proposal was implemented using established biomedical tools and was validated considering a catalogue of datasets related to Alzheimer's disease.

Although this system was applied 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.

Chapter 6

Visualisation of Semantic Data

Medical studies enable a deeper understanding of health conditions, diseases and treatments, helping improve medical care services. In observational studies, it is crucial to select adequate datasets to ensure the study's success and the quality of the results obtained. Biomedical databases often have restricted access policies and governance rules. Thus, an adequate description of their content is essential for researchers who wish to use them to conduct medical research. A strategy for publishing information without disclosing patient-level data is through database fingerprinting and aggregate characterisations. However, this information is still presented in a format that makes it challenging to search, analyse, and decide on the best databases for a domain of study.

Several strategies can be used to visualise and compare the characteristics of multiple biomedical databases. This study focused on a European platform for biomedical data sharing and dissemination. Semantic data visualisation techniques were used to assist in comparing descriptive metadata from several databases. The great advantage lies in streamlining the database selection process, ensuring that sensitive details are not shared. To address this goal, two levels of data visualisation were considered, one characterising a single database and the other involving multiple databases in network-level visualisations.

Regarding observational studies, during the feasibility study phase, one defines inclusion and exclusion criteria and specific database characteristics to construct the cohort. However, the comparison of database characteristics and their evolution over time are not easily identified during this selection. Data comparisons can be made using the data properties and aggregations, but the inclusion of temporal information becomes more complex due to the continuous concepts' evolution over time. Two visualisation methods are proposed to better describe data evolution in clinical registers using biomedical standard vocabularies to overcome this issue.

This study revealed the impact of the proposed visualisations and some open chal-

lenges in representing semantically annotated biomedical datasets. One of this work's outcomes was identifying future directions in this scope.

6.1 Contextualisation

The secondary use of health data is currently a common strategy in medical research to conduct observational studies in various domains (Hripcsak et al., 2015), ranging from pharmacological research to public health policy design (Cheng and Phillips, 2014). Several steps are necessary to plan and execute this type of study successfully. The first step is to define the research question focusing on solving the problems. Therefore, the cohort group is established, with the inclusion and exclusion criteria and the outputs to be evaluated. After formalising the study protocol, researchers must identify relevant information resources and possible data partners. That is accomplished by contacting the database owners to guarantee a data access agreement, analysing the data and publishing the research results (Hripcsak et al., 2016). Along this process, identifying databases of interest is a critical step in conducting high-quality observational studies. Therefore, solutions that simplify searching for biomedical databases may help researchers at this stage.

When assembling a cohort for an observational study, medical researchers need to choose and access the most suitable databases for the purposes pursued. This task becomes simpler when database catalogues oriented to the scientific community's interests to which the researcher belongs are available (Sequeira et al., 2021). Using these resources becomes even more necessary in the case of multicentre studies, where the achievement of the study objectives brings more challenges in terms of research protocol development, work management, and harmonised access to data (Almeida et al., 2021).

Biomedical data is prevalent in the international scientific landscape, enabling the creation of numerous repositories and databases published online (Kolker et al., 2012). For the benefit of researchers who need these sources of information to conduct their research, mechanisms to find data are required. Although retrieving information from web pages using search engines has been usual for many years, the same solution for data repositories is still at an early stage (Brickley et al., 2019). Another way is to use database catalogues. Data owners use catalogues to publish descriptive information about their databases. They provide database fingerprint information, which is characteristics of the database content, including institutional details and the access policies and governance rules (Oliveira et al., 2019; Silva et al., 2018).

Some illustrative examples of catalogues in biomedicine can be considered. Cafe

Variome¹, for instance, enables data discovery based on the semantic similarity of diseases, phenotypes and drugs, relating patient data to the terms of an ontology (Lancaster et al., 2015). YummyData² monitors SPARQL endpoints to collect biomedical linked data (Yamamoto et al., 2018). FAIRsharing³ is another resource that describes and links data policies, repositories and databases, with a strong focus on the natural sciences (Sansone et al., 2019).

Database catalogues make information accessible from a centralised access point. However, sometimes it is still not trivial to choose the best data sources. When there are different databases satisfying different requirements, performing more specific studies may be complex, such as conducting patient-level prediction studies (Bos et al., 2018). In this type of study, researchers need to identify the datasets used to train the prediction models and the datasets used to validate the predictions (Reps et al., 2018). These decisions usually involve a deeper understanding of the datasets publicised by the community. It is desirable to lighten the burden of these choices by having some recommendation mechanism or user-friendly interfaces allowing the data to be queried in a way that is both simple to use and able to provide non-trivial results (Gall et al., 2008).

When using a recommendation system, recommendations are based on rules that can be more or less adaptive to new situations (Almeida et al., 2020). While history-based learning, when using these tools, can lead to better results over time, it is still not the best approach for many cases. Queries can be built using forms that combine the descriptive metadata of the datasets of interest for better results. However, creating more complex queries is not always easy. The use of natural language interfaces can greatly facilitate searches (Pereira et al., 2022), but even with a good search strategy, reading the results may not be intuitive enough.

The presentation of semantic search results for metadata is usually reduced to simple tabular data (Lancaster et al., 2015; Yamamoto et al., 2018; Sansone et al., 2019). However, presenting tabular data does not take advantage of the relationships that link the various entities. Graphs are another common way of giving semantic data (Lopes and Oliveira, 2013), but they may not provide the best information for extensive graphs. On the other hand, mechanisms are also wanted that allows browsing the data and possibly building new queries. Using information visualisation helps improve decision processes. Adding visualisations to biomedical database catalogues allows for better analyse and comparing data from a single database or assessing the network level of several databases.

¹<https://www.cafevariome.org/>

²<http://yummydata.org/>

³<https://fairsharing.org/>

Medical data are constantly evolving (Siegler, 2010). Therefore the time dimension is of interest to the data selection process. Incorporating the concepts' temporal evolution into these two visualisation levels can mitigate possible data scarcity. It is convenient to store information about insertions, deletions, and data changes. Historical data could increase the range of possible options regarding database selection, attending that it is not uncommon for biomedical data catalogues to have a log system allowing for tracking data over time. However, this data is only used to verify the sanity of the solution or carry out data restoration actions in case of system disruptions (Chiueh and Pilania, 2005).

This work explores different visualisation and comparison techniques applied to semantic data. The analysis identifies points that can be improved in a catalogue used to publish metadata from multiple health databases, exemplifying the transverse limitations of the most common catalogues. Possible visualisations for semantic information in different health contexts are shown. The objective is to understand the best strategies to represent data applied to this domain and identify the open challenges for better representation of biomedical datasets. Two types of visualisation that show the temporal evolution of semantic data are also proposed. The first proposal intends to present the data at the database level and allows to see the temporal evolution of a particular selected element. The second proposal aims to visualise the temporal evolution of the data at the semantic network level. The main goal of these visualisations is to improve the use of biomedical data catalogues to help researchers make better data choices for their research studies.

6.2 Background

Visualising and interacting with semantic data improves the way researchers find and perceive the most relevant information for their studies. This section addresses the visualisation and comparison of semantic data and the problem of criteria changing for the inclusion or exclusion of elements in an observational study.

6.2.1 Querying and Visualisation of Semantic Data

The visualisation of semantic data ranges from the simple organisation of semantic triples in tables to the visualisation of graphs taking advantage of the relationships between the different entities. This last form allows richer visualisations; still, it is not uncommon to fall into scenarios where the high number of entities and relationships prevents a clear reading by users. Some examples of solutions for querying and

visualising semantic data can be presented. Yet Another SPARQL GUI (YASGUI)⁴ (Figure 6.1) is a SPARQL client that uses module tabs to allow independent access to multiple endpoints (Rietveld and Hoekstra, 2013). The tool is packaged with auto-complete support, syntax checking, syntax highlighting, query sharing, query retention, and file upload/download.

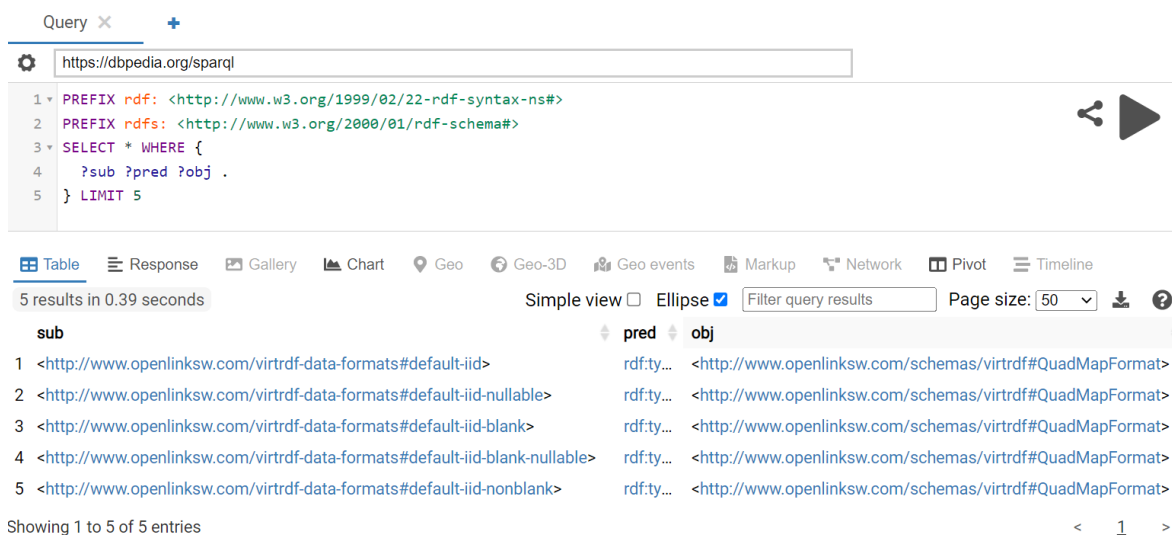


Figure 6.1: Screenshot of the YASGUI interface. The SPARQL queries input area can be seen at the top of the available module tab. At the bottom pane, can be seen the query results using different views.

SPARQLGraph⁵ is a web-based platform implemented using the diagramming library mxGraph⁶ for graphically querying biological semantic databases (Schweiger et al., 2014). Users can compose graph queries on a drawing board by adding new visual elements (nodes and edges). Users can only choose between elements resulting from a previous choice made by the tool’s creators, which is a limitation.

The PIBAS FedSPARQL⁷ (Djokic-Petrovic et al., 2017) solution (Figure 6.2) was applied to a use case where data is collected from tests with bioactive substances and annotated against an ontology. The proposed solution enables the federation of those data with supplementary information that can be extracted from global initiatives such as Bio2RDF (Callahan et al., 2013), Chem2Bio2RDF (Chen et al., 2010), and the EMBL-EBI platform (Li et al., 2015). The system offers templates and generates static federated SPARQL queries for retrieval of relevant information. The results are presented in tabular form.

The Spatial-Temporal Content Explorer (SPEX)⁸ (Scheider et al., 2017) is a tool for

⁴<https://github.com/TriplyDB/Yasgui>

⁵<https://github.com/tadKeys/sparqlgraph>

⁶<https://jgraph.github.io/mxgraph/>

⁷<https://github.com/marijadjokic/PIBASFedSPARQL>

⁸<https://github.com/lodum/SPEX>

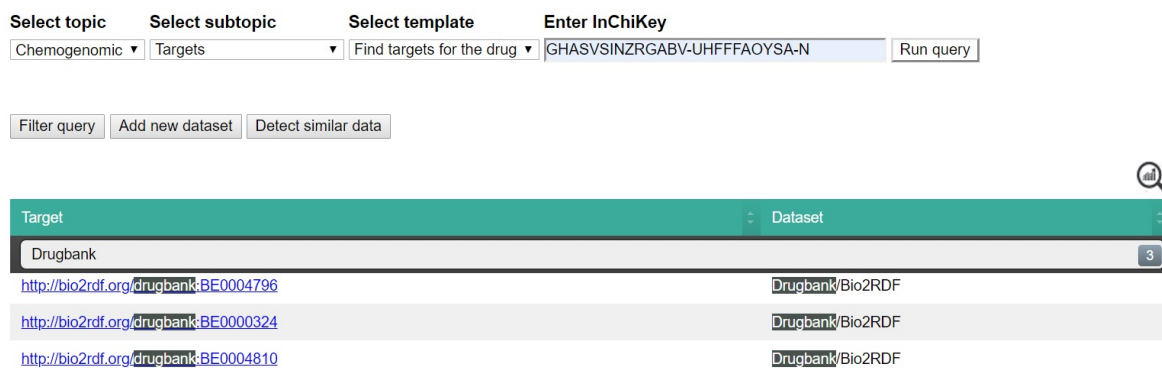


Figure 6.2: Screenshot of the PIBAS FedSPARQL interface. At the bottom, can be appreciated the tabular view of the results of a query.

visualising the temporal and spatial dimensions encoded in semantic data. Users can construct queries using a graph, or they can issue SPARQL queries directly. Figure 6.3 shows the different panels of interest in the application interface.

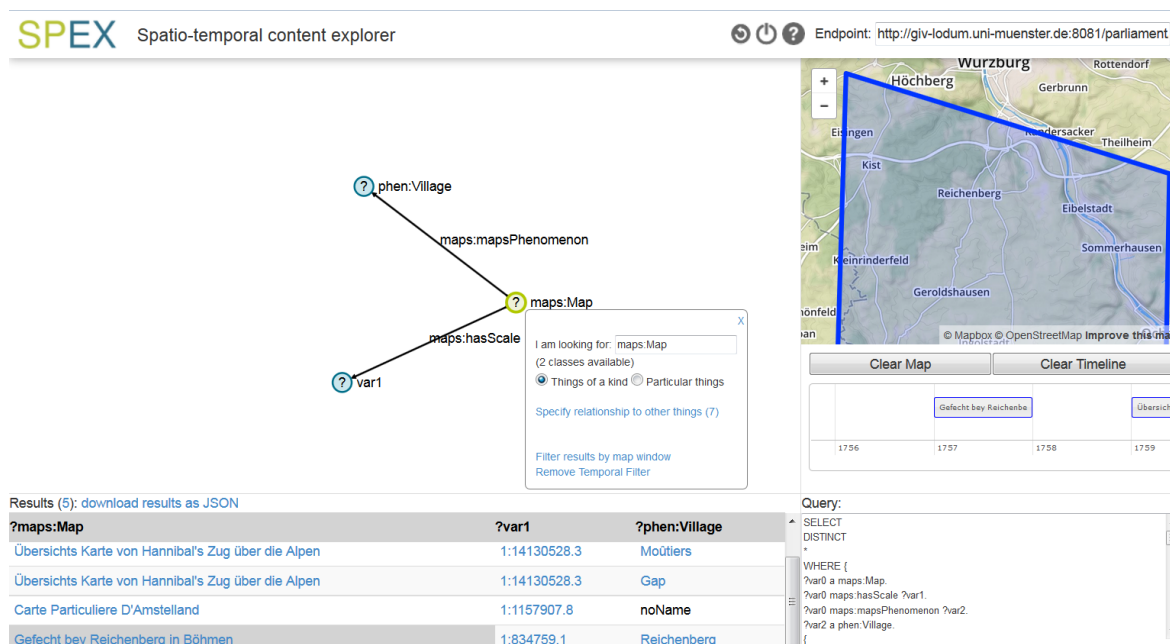


Figure 6.3: Screenshot of the SPEX interface. At the upper left is the query pane to construct query patterns. The space-time filter pane is at the upper right. At the bottom is a tabular presentation of results and a SPARQL query box.

Lekschas and Gehlenborg (2017) proposed SATORI⁹ (Figure 6.4), an integrative search and visual exploration interface for biomedical data repositories. It allows performing ontology-guided visual exploration, enabling researchers to search, browse and semantically query data repositories seamlessly. The solution is based on a fixed list of datasets and does not automatically incorporate a methodology to infer structural

⁹<https://satori.lekschas.de/>

information (ontology). Nor can be connected to an arbitrary SPARQL endpoint, immediately starting to navigate the data.

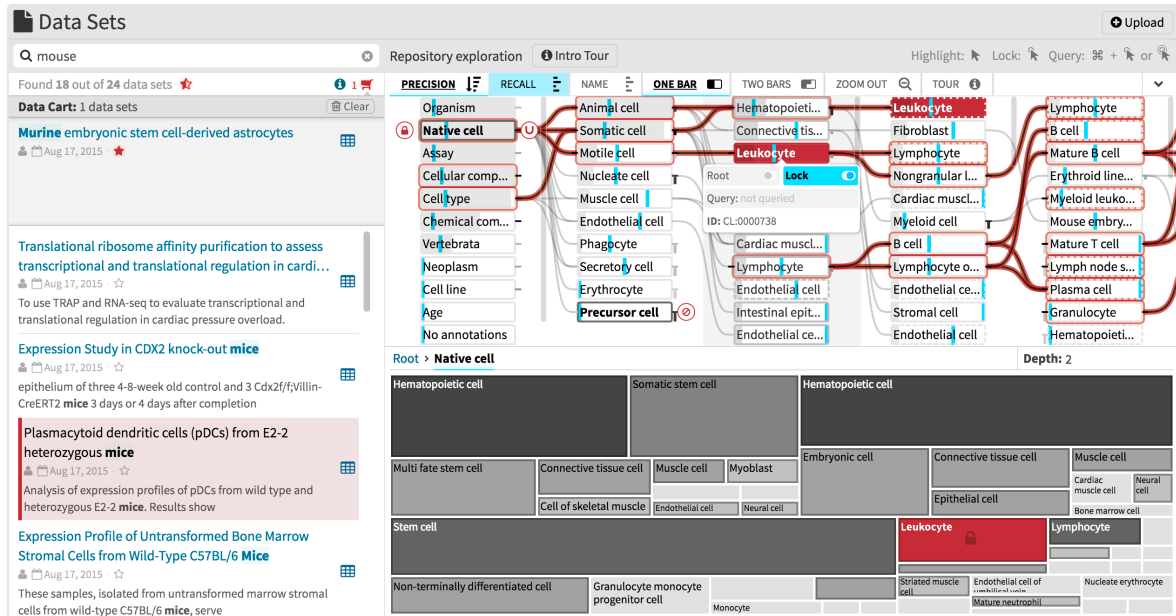


Figure 6.4: The interface of SATORI consists of the dataset and exploration view.

Semantic data have a graph or network structure that allows graph visualisations, emphasising the relationships between the various entities. Elaborated and high-level programming languages with abstraction layers can be directly used to construct graphs, establishing a balance between expressiveness and ease of programming. The lower the degree of abstraction, the greater the requirement for programming knowledge and the lower the productivity. Some data subtleties can be captured with lower-level programming, which otherwise might go unnoticed. However, visualisation libraries are naturally used in most applications because they provide convenient resources for various applications (Li et al., 2022).

Despite the various display options presented, limitations remain. There are strategies to create different dashboards to visualise data and compare them. However, regarding semantic data, the availability of filters is more limited. For example, when a researcher wants to select a set of databases to answer the research questions and notices that the initial cohort is not the most suitable for the study to be carried out. Or when adjusting some of the returned parameters according to specific needs, such as adding a certain threshold. Another typical situation is when choosing a set of patients suffering from a particular disease or taking a specific drug. It is necessary to navigate the underlying ontology by visually adjusting the parameters to generate new data queries.

6.2.2 Interacting with Semantic Data Visualisations

A visual query system (VQS) is a non-formal solution to database querying that uses visual representations to depict the domain of interest and help express knowledge base queries. VQS divides into form-based, diagram-based, icon-based, faceted, and hybrid. A form (e.g., a table) is a named collection of objects with the same structure and is the most basic approach after plain text use. Diagram-based representations (e.g., a graph) allow exploring and showing relationships between entities. Icons denoting entities allow performing queries by combinations according to some spatial syntax on icon-based solutions. Faceted systems use classifications that organise items into multiple independent views. Solutions that combine more than one visual representation are called hybrid solutions (Catarci et al., 1997; Lloret-Gazo, 2016).

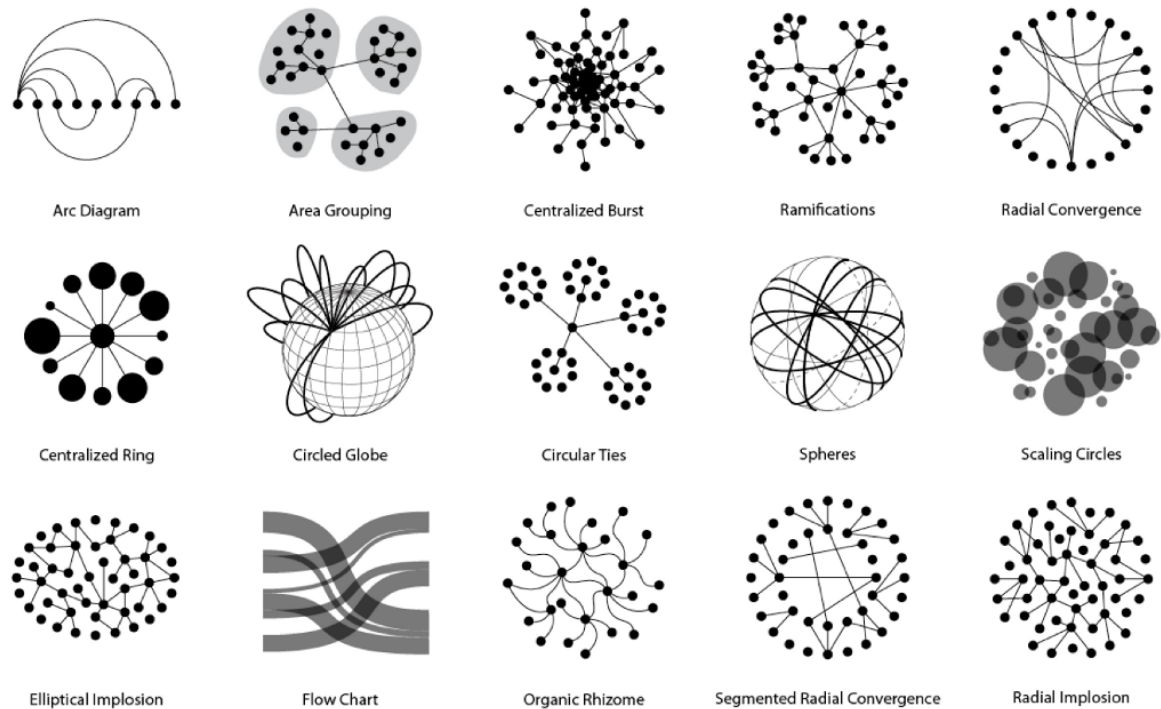
Proper development of information systems depends on understanding users' needs. It is necessary to have exploratory search mechanisms to use data repositories helpfully. As highlighted by Marchionini (2006), short queries typed into search boxes do not fulfil current users' needs. Compared to analytical search strategies that depend on a carefully planned series of questions, browsing depends on on-the-fly choices, encompassing selection, navigation, and most importantly, trial-and-error tactics. Query suggestions are also expected. This dynamic behaviour poses many challenges because the system should follow users' expectations, providing rich interactive features.

Strictly related to user interfaces allowing working at multiple levels of detail, Cockburn et al. (2009) identified four approaches: overview+detail, zooming, focus+context, and cue-based techniques. The overview+detail approach creates a spatial separation between contextual and detailed information. Zooming allows perspectives with varying degrees of proximity to the objects of interest (with temporal separation between the views). Focus+context allows integrating both focus and context into a single display. Finally, cue-based depicts the elements modified to highlight, suppress, or contextualise them.

In general, information visualisation has two main dimensions: representation and interaction. Regarding the first dimension, there are many possibilities for graph representation, as can be seen in Figure 6.5, like arc diagrams, area grouping, centralised burst, radial convergence, centralised ring, circled globes, circular ties, spheres, and scaling circles (Lima, 2011).

Yi et al. (2007) and Heer and Shneiderman (2012) pointed out the following guidelines for interaction techniques:

- Visualise data by choosing visual encodings;
- Organise multiple windows and workspaces;



Source: Lima (2011).

Figure 6.5: The 15 typologies of network visualisation. In arc diagrams, a single axis is used to display all nodes and semicircles to represent the arcs. Area grouping makes clusters of interconnected nodes evident, and a centralised burst highlights important nodes identifiable as highly connected. Tree graphs can be visualised using ramifications. Radial convergence allows visualising relations between nodes arranged in a circle. A centralised ring can be used to check the relationship of multiple nodes with a single central node. Circled globes are projections of other topologies on a globe. Circular ties connect several centralised rings to a central node. It can be helpful to have nodes and arcs drawn on spheres. Scaling circles allow aggregating of similar nodes. Other basic network depictions include elliptical implosion, flow chart, organic rhizome, segmented radial convergence, and radial implosion.

- Select and mark something as interesting;
- Explore (navigate) to show something else;
- Reconfigure to deliver a different arrangement;
- Encode to offer a diverse depiction;
- Abstract (elaborate) to see more or less detail;
- Filter to see something conforming to a condition;
- Connect (coordinate) to see related items
- Sort items to expose patterns;
- Derive values or models from source data;
- Record analytics history for revisitation, review, and sharing;

- Annotate patterns to document findings.

Interaction must adapt to the particular problem to be solved, and it organises around a user’s intent, hiding the system’s low-level interaction details. In our context, visualisation means choosing the application layout that best fits users’ intents. It ties in closely with organising multiple windows and workspaces. The “select” feature allows users to mark and track items of interest, like nodes and edges. When exploring (navigating), users want to examine a different subset of data cases to gain understanding and insight. The abstraction/elaboration interaction provides the necessary level adjustment of a data representation, while the filtering can reduce the representation’s complexity by hiding the elements that are not relevant to the user. The “connect” primitive traces the same object when presented simultaneously in different views. The sorting operation is used to surface trends or organise data around some analysis unit. Through their actions, users create imminently unrepeatable hypotheses and generate chains of queries the app must save for future use. The “record” and “annotate” features are helpful to deal with that issue. “Encoding” allows changes in the visual appearance of each data element, like changing size or shape. In a more general way, the reconfiguration feature must provide users with different perspectives to uncover hidden characteristics of nodes and their relations.

Users’ behaviour is iterative and depends on their cognitive style (Knight and Spink, 2008). However, when referring to usability, aspects belonging to the domain of behavioural sciences are not relevant in contrast to the scientific understanding of usability based on experimental data. Evaluation is essential to assess the system’s relative success compared with others (Elmqvist and Yi, 2015). Elbedweihy et al. (2015) overviewed semantic search evaluation initiatives, pointing out the importance of considering information retrieval evaluation activities in general. It is interesting to know how users’ search requests are handled by performing a system-oriented evaluation. Equally crucial are user-oriented assessments. Efficiency, learnability, utility, and user satisfaction can be highlighted. Typical assessment tools include event logs, think-aloud, and questionnaires. As a final note, it can be mentioned that Hilbert and Redmiles (2000) extensively studied the extraction of usability information from user interface events by processing logs.

Queries can be intuitively built using a query builder with visual artefacts. However, it is necessary to go further and obtain visualisations that allow reissue questions and manipulate the results in a flexible way that will enable comparisons and refinements (e.g., to redefine cohorts).

6.2.3 Time-evolving Semantic Data

Temporal data management allows querying, accessing and navigating through different data versions to understand their evolution or choose pieces of information from a given moment that are more suited to the user's interests (Kaufmann et al., 2013). Within the scope of relational databases, the temporal dimension is considered by creating specialised data structures to optimise accesses. The same principle is valid for pure graph databases. For instance, Khurana and Deshpande (2016) proposed a historical graph store for large-scale volumes of data integrating a new temporal graph index and a temporal graph analysis framework to perform complex temporal analytical tasks.

Time coding strategy can be divided into copy systems or log systems (Böhlen et al., 2017). With each change, the updated full copy of the data is saved in the copy approach. In the log approach, the first complete version of the data is kept, and changes are recorded in a log. Hybrid systems that consider both approaches can also be adopted.

Querying data that evolves can follow alternative patterns (Salzberg and Tsotras, 1999). The first considers a time interval and extracts valid entities for that time interval. Another querying approach takes a time interval and a set of entities to retrieve those entities' temporal evolution. Finally, one can take just a collection of entities and check their entire evolutionary history.

The visualisation of semantic data considering the temporal dimension allows observing patterns and determining when there is a greater concentration of entities of interest. Time-oriented data visualisation techniques can be classified from the arrangement point of view as linear or cyclic and from the time primitives point of view as instant oriented or interval (Aigner et al., 2011).

The discovery and study of patterns are facilitated when using time curves, violating the linearity of the spatial provision of the most usual timelines. Bach et al. (2016) consider a non-linear time tape curving according to data similarity at each moment, with the advantage of being possible to ascertain the depth of the changes.

6.3 Databases for Observational Health

Observational Health Data Sciences and Informatics (OHDSI)¹⁰ initiative (Hripcsak et al., 2015) is an international, interdisciplinary, multi-stakeholder project to develop applications to access and analyse large-scale observational health data. The core of this project lies in adopting a common data model for the treatment of health data.

¹⁰<https://www.ohdsi.org/>

Solutions to extract, transform, and load data from different sources in the proposed standard format are available, as well as other tools related to the data modelling process.

Observational Medical Outcomes Partnership (OMOP) promotes the proper use of observational healthcare databases (Stang et al., 2010). OMOP Common Data Model (CDM)¹¹ has been proposed as an open relational data model standard designed to establish the structure and content of observational health data. OMOP CDM allows the creation of relational databases to load transformed data from other sources of information. In this data schema, a set of tables was defined to store the standard vocabularies in an interoperable structure. These tables can represent each vocabulary and all the information associated with it. This is essential to ensure database interoperability in multicentre studies when using institutions that adopt different vocabularies in the original data. The tables are defined in the collection denominated “Standardised Vocabularies”.

ATHENA¹² (which stands for “Automated Terminology Harmonisation, Extraction and Normalisation for Analytics) is a standard vocabulary repository based on an automated building process. ATHENA allows keyword searching for terms using filters to select the application domain (drugs, conditions, procedures, devices, observations, and measurements), type of concepts (for classification, standard or not), class, vocabulary and validity. The search results are presented in a tabular format, and it is possible to browse the terms shown for those lying inside a hierarchy.

6.4 The EMIF Catalogue Use Case

The European Medical Information Framework (EMIF)¹³ initiative focused on creating a European Medical Information Framework to provide better healthcare using the vast amounts of biomedical data available. A web platform was thus designed to offer a database catalogue (the EMIF Catalogue¹⁴) where data custodians can publish metadata about their biomedical databases with different levels of granularity. The EMIF Catalogue also enables the creation of communities that gather around common interests and that, in this way, share and have access to biomedical data of interest to them (Oliveira et al., 2019).

For each database described in the catalogue, the data custodian must provide information that constitutes the database fingerprint (Figure 6.6). This information

¹¹<https://ohdsi.github.io/CommonDataModel/cdm54.html>

¹²<https://athena.ohdsi.org/>

¹³<http://www.emif.eu>

¹⁴<https://emif-catalogue.eu/>

has several fields, such as the name of the database, identification of the institution that owns it, its location, and the person in charge, among others that were defined collaboratively by the community members. The fingerprint also contains data relating to the database content, like the number of subjects and clinical information. Therefore, researchers can find the databases relevant to their investigation by consulting these fingerprints.

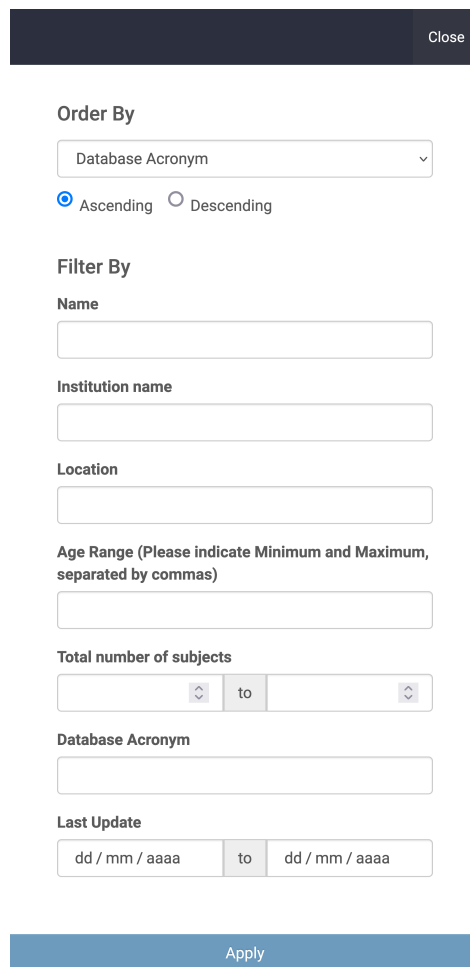
Figure 6.6: EMIF Catalogue database questionnaire form to collect database characteristics. On the left are several fields to be filled in, and on the right, the different categories of data to be entered together with the current filling status.

6.4.1 Searching and Visualisation Features

Searching features over biomedical catalogues is common among medical researchers to identify databases of interest. There are currently several alternatives for searching for biomedical databases in the most common catalogues, and in particular, in the EMIF Catalogue. A basic search is to filter the substrings of a word. For example, a researcher who searches a database with records of patients with Alzheimer’s disease and starts by entering “alz” will see the system’s suggestions. Another form of basic search is the selection of value windows, considering a lower and upper limit.

For structured searches, a simple form can be used with fields operated by the logical conjunction operation (Figure 6.7). These forms are composed of the most relevant concepts of the fingerprints collected in each community. The main problem with this approach is that not all fingerprint concepts are considered.

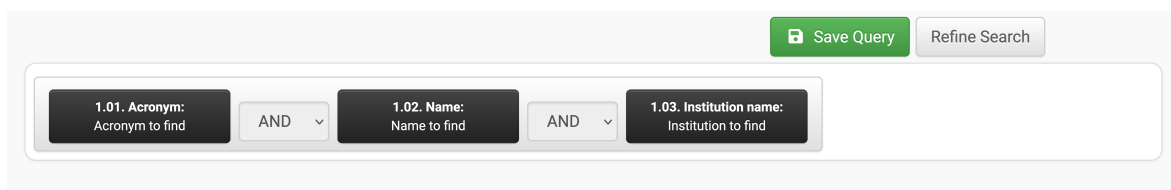
To define a filter using the remaining concepts, EMIF Catalogue has a different query builder. For more complex questions, a researcher can use a form combining all



The image shows a simple query form for the EMIF Catalogue. It features a dark blue header bar with a 'Close' button on the right. Below the header, the form is organized into sections: 'Order By' with a dropdown menu set to 'Database Acronym' and radio buttons for 'Ascending' (selected) and 'Descending'; 'Filter By' with several input fields for 'Name', 'Institution name', 'Location', and 'Age Range (Please indicate Minimum and Maximum, separated by commas)'; a 'Total number of subjects' section with two dropdown menus and a 'to' separator; a 'Database Acronym' input field; and a 'Last Update' section with two date input fields in 'dd / mm / aaaa' format. A blue 'Apply' button is located at the bottom of the form.

Figure 6.7: EMIF Catalogue simple query form.

possible options and disjunction operators in addition to the conjunction (Figure 6.8). The results obtained after defining such filters are presented in a result list.



The image shows an advanced query form for the EMIF Catalogue. It includes a 'Save Query' button (green) and a 'Refine Search' button (grey) at the top right. The main query area contains three filter conditions: '1.01. Acronym: Acronym to find', '1.02. Name: Name to find', and '1.03. Institution name: Institution to find'. Each condition is connected to the next by an 'AND' operator with a dropdown arrow.

Figure 6.8: EMIF Catalogue advanced query form.

The platform also allows comparing a small set of databases against a reference database (Figure 6.9). Although this type of comparison is successful in some scenarios, it lacks an overview of the database network in the health domain.

	DDA	DDC	DDB
Contact Details			
Database Acronym	DDA	DDC	DDB
Database Name	Demo Database A	Demo Database C	Demo Database B
Institution name	University of Aveiro	Polytechnic Institute of Bragança	University of A Coruña
Department name	IEETA	CeDRI	DICT
Administrative Contact			
Scientific Contact			

Figure 6.9: EMIF Catalogue databases comparison view.

EMIF Catalogue also allows selecting the question sets and databases to be exported to a spreadsheet. In this case, the researcher can use the Excel features to navigate the data, which is not user-friendly since it requires a third-party tool. The view to define this filter is a two-column list to select databases (Figure 6.10).

6.4.2 Steps for Improved Biomedical Metadata Visualisation

The EMIF Catalogue is a platform for biomedical data discovery that adheres to FAIR (Findable, Accessible, Interoperable, and Reusable) principles (Trifan and Oliveira, 2018). The solution supports data sharing for different communities, such as the community interested in research on Alzheimer’s disease. The system uses ontologies to model the metadata to allow the annotation of several levels of information to describe the databases registered in the catalogue. The community members can annotate the

Figure 6.10: EMIF Catalogue two-column list selector.

concepts for the questions, and at a deeper level, they can annotate the answers to the question in the questionnaire (fingerprint). They can also have higher annotations, namely to the community itself, so the community can be related to others in the system that share the same interests.

Although the initial analysis focused on the EMIF Catalogue, it was noticed that the tabular format is the most commonly used in such platforms. Some may have charts representing specific concepts, but a lack of visualisations using semantic data in health database catalogues is identifiable. This fact significantly limits users' options when selecting the databases of interest for a new research study, resulting in the reuse of databases that the researchers are familiar with instead of selecting others in the community with the potential to empower their findings.

6.4.3 Measuring User Behaviour

Quality improvement is part of the software lifecycle and can be guided by metrics and thresholds that trigger the improvement process (Agnihotri and Chug, 2021). Evaluating the information visualisations currently available in the catalogue can show how best to introduce improvements. The platform has more than 1,600 registered users, distributed among 11 communities. The system was initially designed to collect just a few metrics about specific views for debugging and functionality improvements. However, this log is helpful in providing an overview of user actions on the platform. Only the records of the last two years were used for this study.

EMIF Catalogue uses in its core the Django-Hitcount¹⁵ for collecting user metrics. This package counts the number of hits for a particular object in the code. For instance, the number of hits on buttons and links that open the different views that expose the information searched for by users was studied by looking at the logs.

From the study of the logs, the conclusion is that users prefer to use views that allow individual dataset fingerprints to be consulted. However, the lack of comparison features in this platform and the appeals in some communities for overviews of the databases in the network motivated this work.

6.5 Ontology-driven Visualisations Scenarios

Visualising semantic data gives a perception of different situations and guides users in decision-making. Decisions can be based on studying database descriptions, comparing databases, and filtering and browsing data. These three search levels allow for

¹⁵<https://django-hitcount.readthedocs.io/en/latest/>

informed choices in conducting observational studies. It is also necessary to consider adequate data structures to capture the temporal evolution of concepts. In this section, visualisation proposals for all the aforementioned levels of abstraction are presented.

6.5.1 Temporal Knowledge Bases

It is necessary to define a data structure to capture the temporal evolution of the entities and relationships of an ontology. Recalling the knowledge base concept:

A Knowledge Base is an edge labelled multi-digraph $K = (V, E^)$ that is defined by a node set $V = V_1 \cup V_2$ and a labelled arc set $E^* = \{(v_1, l, v_2) : v_1 \in V_1, v_2 \in V_2, l \in L\}$, l being an element of the label set L .*

Adding the time dimension to this data model allows the following definition:

A Temporal Knowledge Base (TKB) is a triple of the form $K = (V, E^, T)$, with V and E^* as defined before, and a set $T = T_i \times T_f$ of timestamps.*

The ontological concepts and individuals constitute the set of vertices. Two timestamps are associated with each entity. The first timestamp, $t_i \in T_i$, records the moment of inserting the element in the KB. The second timestamp, $t_f \in T_f$ reports the moment of concept evolution (removal or alteration). Thus, a timespan is implicitly defined, useful for applications, namely when discussing visualisations.

Following FAIR principles, removing a concept does not determine its exclusion from the TKB. The TKB arcs indicate the relationships between the different KB entities. As a simplification, the temporal dimension is considered only for concepts, thus excluding individuals. The relationships that make up the KB arcs are also temporally annotated. Relationships are only marked when they are inserted to avoid inconsistencies. Thus, the evolution of a relationship generates a new relationship that does not affect the triples previously entered.

6.5.2 Database-level Visualisations

Information visualisation makes it easier to choose whether to include or exclude a database when faced with many database descriptors. Treemaps provide a visualisation of data hierarchies using nested coloured rectangular shapes. Treemaps are an alternative to visualise hierarchical structures in a compact way that allows a quick view of the relationship between the amounts of elements for each data category. When creating this type of visualisation, each category is assigned a rectangle subdivided into smaller nested rectangles representing the subcategories of data. Each rectangle size

is calculated by taking the proportion of elements from each category or subcategories concerning all data. It is usual to use different colours to allow even easier reading at a glance. This type of visualisation is not suitable for ontologies that contain cycles, as it generates a recursion phenomenon that prevents the construction of the treemap.

The EMIF Catalogue has thousands of instances for a wide range of concepts. In addition to viewing how item percentages affect the size of rectangles, the treemap view must provide numerical information (Figure 6.11). When selecting one of the rectangles, more detailed information about that entity should be presented.

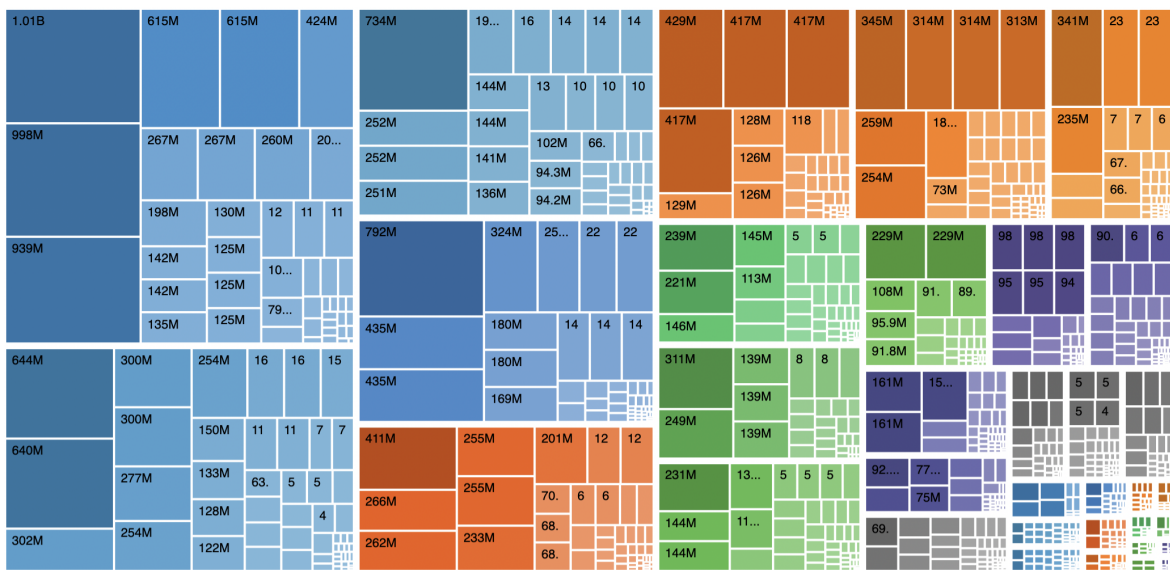


Figure 6.11: UI mockup proposal of a treemap visualisation.

Semantic data find the most natural form of representation in visualisations that use graphs since it is easy to appreciate the entities and the relationships between them (Dadzie and Pietriga, 2016). However, when the number of elements increases, there is a significant loss of legibility of the presented information. The strategy to address this problem is to focus on some criterion that allows the creation of more understandable visualisations for users. Of the multiple possible criteria, the interest in visualisation at the database level aims to focus the user's attention on that database. Therefore, the entities and relationships related to this information must appear prominently in the foreground, introducing a differentiation that can be achieved by changing the visualised elements' dimension and colour.

Figure 6.12 shows a graph representing links between different EMIF Catalogue concepts. Each entity is represented by a point that can be clicked to obtain more detailed information. Researchers should be able to navigate the graph by selecting successive points. In this way, it is possible to perceive how the different represented instances are related.

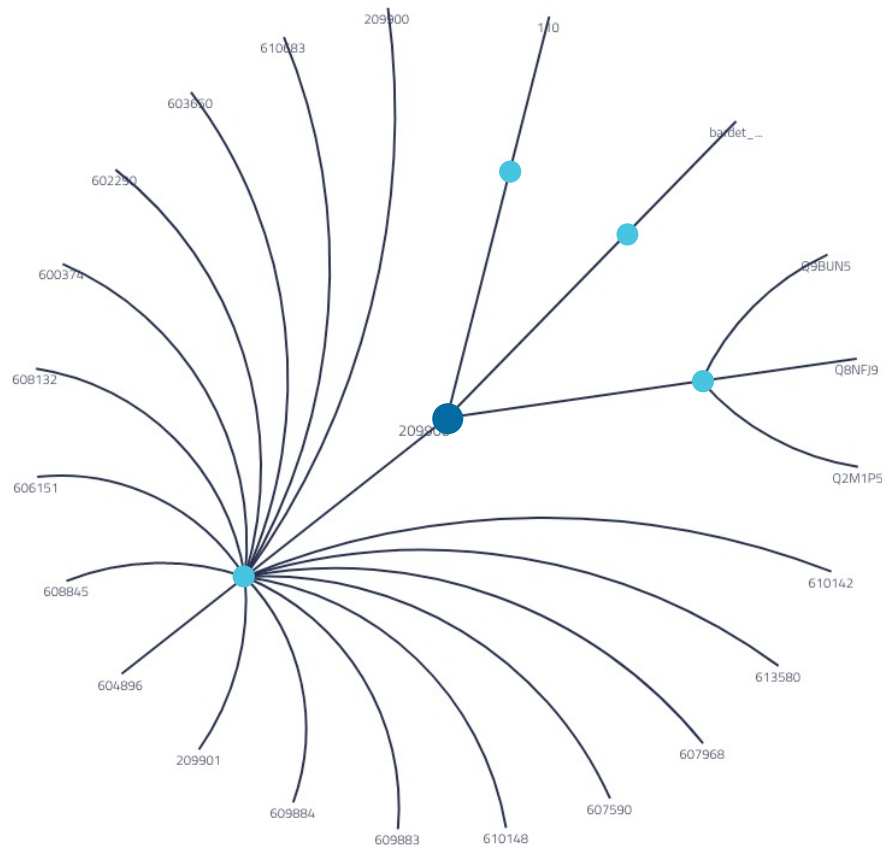


Figure 6.12: UI mockup proposal of graph visualisation.

The capture of the temporal dimension allows the study of the evolution of data classes and their instances. This leads to the need to store the history of the elements to be monitored on the data side. The storage functionality is already implemented in the EMIF Catalogue, which keeps historical data in log files that can be accessed to operationalise the temporal visualisation of the entities of interest. From the visualisation point of view, a timeline is available for each entity or relationship clicked on by the user. There is also a snapshot of a given instant where researchers can see data at that given point in time.

For the EMIF Catalogue, there is interest in a graph-type visualisation in which it is possible to choose any node and see its temporal evolution highlighted (Figure 6.13). If there is no associated historic data, only the current instance should be presented. The timeline allows navigating through different moments in time to study the state of the selected entity.

As new concepts are added, modified or removed from the ontology, the different versions that document these changes are saved and serve as a basis for visualisations. Visualisations that do not consider other moments in time aggregate all the information, making the presentation of different concepts confusing. With the proposed

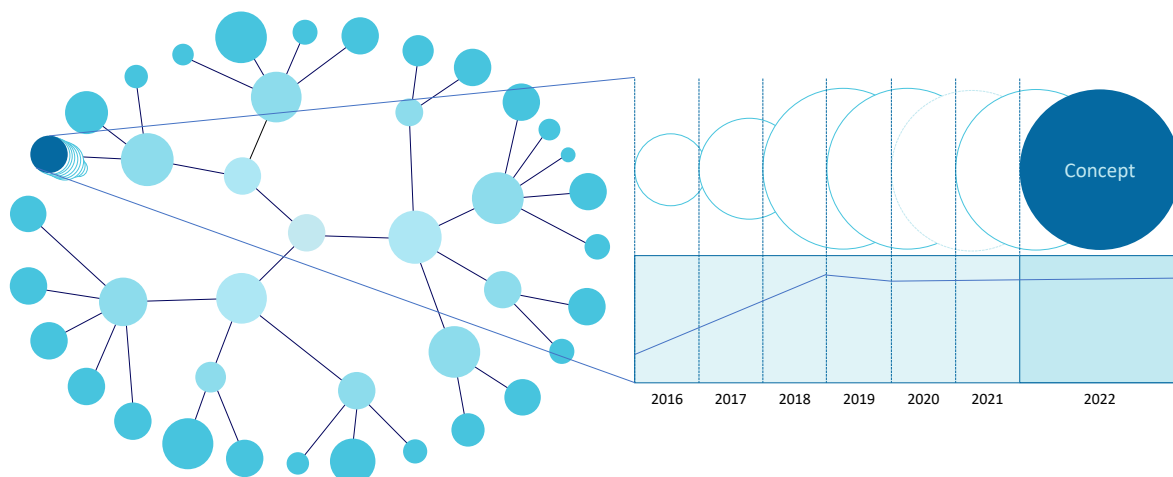


Figure 6.13: UI mockup proposal of a temporal chart visualisation (entity-level view).

visualisation, one can navigate the graph and discover temporal details for each node individually, which helps to have a clearer perception of the data.

6.5.3 Network-level Visualisations

Comparing databases allows informed choices about the data to use in medical studies. The possibility of visually comparing the contents of different data sources is an added value. Using a tabular view, researchers can inspect row by row or column by column to see the greater or lesser density of the data, that is, to understand how many records of each concept there are in each database. However, when using tables, as is currently done in database catalogues similar to the EMIF Catalogue, it is challenging to identify the databases of interest. Depending on the study, sometimes it is necessary to identify a database to be used for analysis and others for validation. An example of these cases is the patient-level prediction studies, in which one database is usually chosen to train machine learning models. These are tested and validated using other databases. Knowing the number of samples for the concepts in the study helps determine which databases should be used for training the models. The different ways of comparing semantic data related to various datasets start from graph views in an attempt to find comparisons and hierarchies.

A strategy to compare databases is to highlight hierarchical relationships extracted from the metadata. This form of structuring can be assumed when defining the ontology. In fact, “is-a” and “SubClassOf” relationships allow obtaining dendrogram representations allowing navigation from a root node to the various branches and leaves. The right side of Figure 6.14 shows the dendrogram that results from processing data from the matrix presented on the left side of the figure. The ontology level corresponds

to a central node that defines its identification and highlights the ontology’s hierarchical structure. The concepts are usually under this node, but an intermediate level (database level) was added at the network level, which allows connecting the existent concepts in each database under this ontology node. The concept level can have multiple layers. However, this idea was simplified by presenting only the leaf nodes. By simply inspecting the degree of intensity of the colour of the leaf nodes at the concept level, can be seen which databases have more elements of a given concept. Arcs model the relationships between the different nodes. For example, a quick inspection of the dendrogram connections shows that concept 4553810 exists in databases A and B.

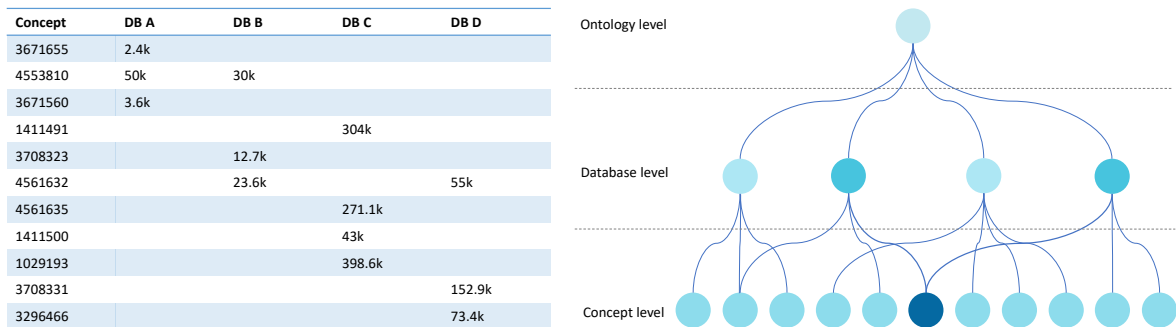


Figure 6.14: UI mockup proposal of a dendrogram visualisation.

Although filters omission in this representative figure, the researchers need to search by the concepts or domains they want to use in the study. The dendrogram is then updated based on the filter applied. Several sub-layers of relations may appear between the database and concept level layers, depending on the searched concepts. Only some nodes at the concept level are presented to avoid overloading the visualisation. It is possible to focus attention on particular concepts, choosing a node and activating it to navigate to related nodes, namely in cases where the same concept has different identifiers in different vocabularies.

There are concepts with an enormous diversity of child concepts, so it is possible to obtain more details visually using a view like these. For example, in the SNOMED vocabulary (Stearns et al., 2001), the concept “Aspirin” has more than 700 child concepts when considering the relation “Specific active ingredient of”. The table supports the dendrogram because the profusion of child concepts makes the visualisation more challenging to use due to the number of leaf nodes.

Graphs are very effective for comparing concepts in different databases. Although these are good to represent entities and relationships of each database, they can also add other linked information to the visualisation that can be semantically asserted. Besides this visualisation possibility, each node or edge represented can be selected to obtain more information. As the amount of data to be presented can increase, this type of

visualisation provides layers that minimise information overload and enhance the focus on relevant information. The representation of these data is similar to Figure 6.12, with the addition of one extra layer in the hierarchy of the ontology corresponding to each database.

Semantic visualisations can use non-traditional formats for specific domains, using images in their composition. This technique can be used when the elements of a given variable have a visual translation that simplifies the representation of the information and increases its value. An example of applying the technique is using drawings of the geographical representation of countries. In this way, one obtains illustrations of the geographic origin of the data that are much easier to grasp than merely reading values in tables. Figure 6.15 shows different databases from different European countries. The general scenario with all countries and respective databases simultaneously and with the same detail can be seen on the left side of the figure. On the right, by selecting some of the countries (e.g., Portugal and Spain), can be seen highlighted the nodes that represent the databases of these countries.

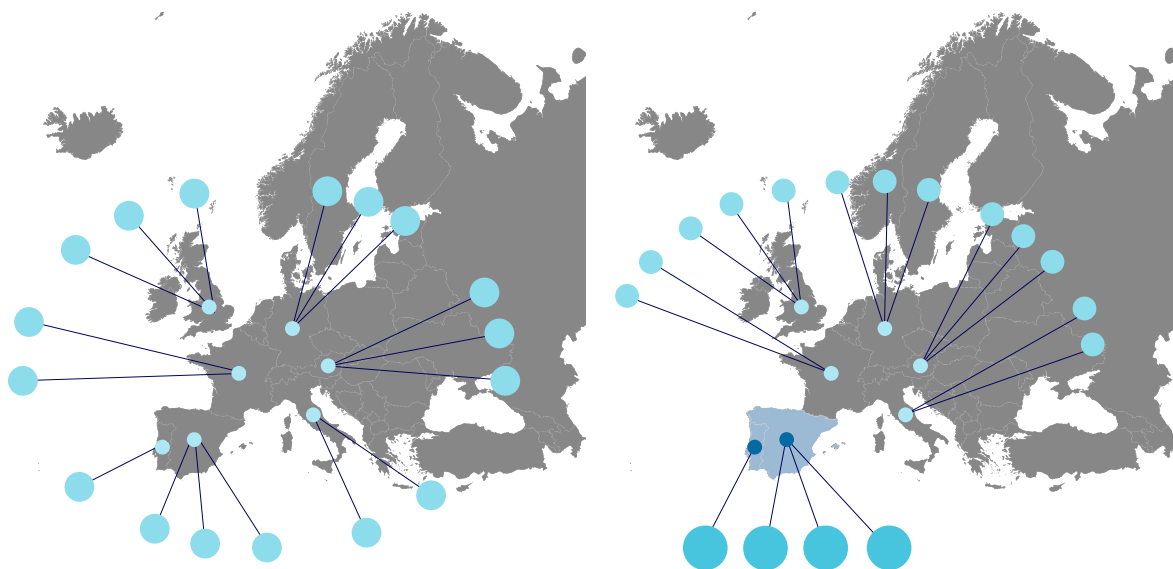


Figure 6.15: UI mockup proposal of map charts visualisation.

Counting distinct database records allows an understanding of whether a particular choice will provide adequate data to conduct a study. Heat maps are a quick and condensed strategy to understand which databases concentrate more data on a given variable (e.g., number of patients) than previous visualisations. This type of visualisation takes two dimensions and expresses the magnitude of a given variable by gradually varying the colour. For a variable with few registers, a paler colour tone is used, and for having many more, a loaded one is used. This graphic representation enables a quick understanding of which databases concentrate more records related to a given variable.

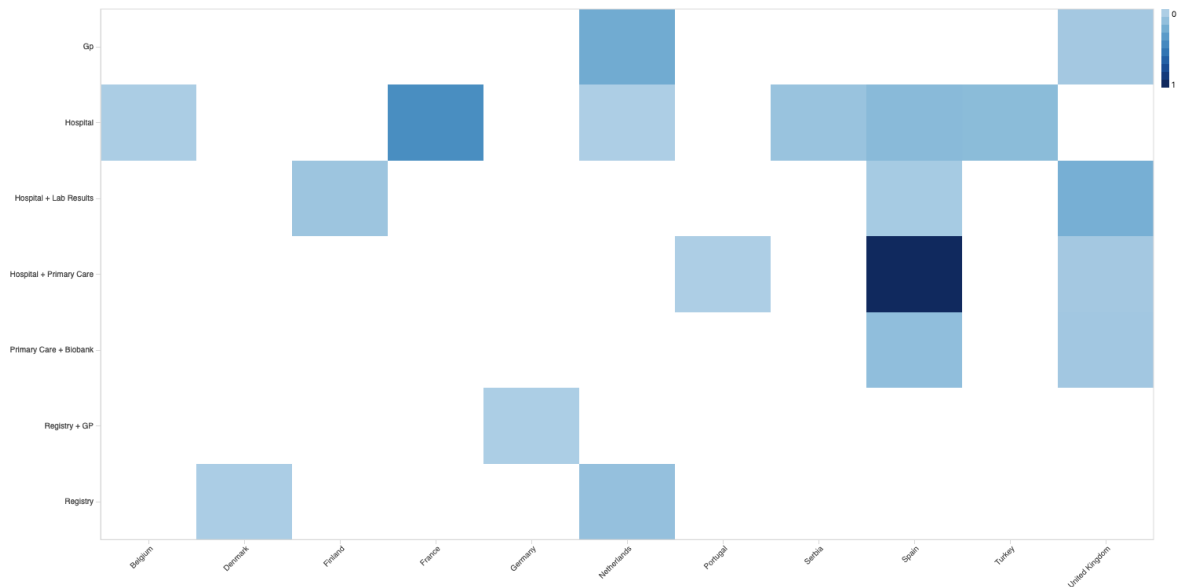


Figure 6.16: UI mockup proposal of a heat map visualisation.

Figure 6.16 shows an example of simplifying the cross-reference information from databases relating to different domains and different countries of origin of the data. Darker colour tones indicate the existence of a more significant number of records.

6.5.4 View Refinements

One crucial step when choosing the databases of interest to conduct a study is the stage in which the researchers need to understand the study feasibility regarding the study protocol and data available. To maximise the study’s success, this step may require several refinements since one of the main issues in medical studies is the lack of subjects with characteristics compliant with the study needs (Rosenbaum, 2017). Analysing some aspects of semantic information representation more deeply allows for gaining perspective on details that might otherwise go unnoticed. The most basic way of manipulating a graphical representation of information is by zooming in on specific details of a graph.

The refinement of the information displayed can be achieved by combining it with a form for selecting values. The side-by-side view of the value refinement form and the preview pane is a powerful tool to guide users’ choices. When a node of interest is selected, researchers can make choices from a range of values and observe the impact of this choice on the visualisation being presented. In this way, they create new queries to the data and get the results interactively.

Visualising data with a temporal dimension should allow smooth navigation between moments in time. The chart must offer a timeline for each selectable visual element

and all elements being viewed. This does not prevent the existence of static elements in time, that is, elements without historical values. In addition to the timeline, each element with history must present some overlapping representation that indicates the trend of evolution of values in the window of the closest past and future times, when applicable.

Figure 6.17 presents a graph-level representation with a temporal dimension. The bottom contains a timeline that allows selecting a particular year and seeing the state of the data network at that moment, as shown at the top of the figure. This feature helps a researcher navigate the different versions of database characteristics and understand the evolution of specific concepts over time, which may influence the selection of the database for the study.

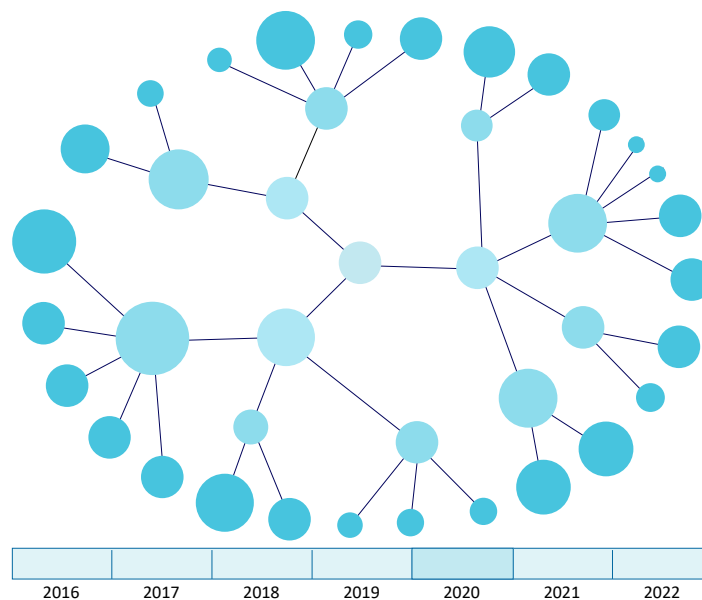


Figure 6.17: UI mockup of a temporal chart visualisation (graph-level view).

One can interact with a visualisation that figuratively represents some of the semantic information nodes based on the position of these elements. The selection actions must drive a reconfiguration of the information presented, giving greater focus to the parts linked to the selected visual artefacts. It is not desirable that the components that pass into the background disappear from view, but rather that they become less prominent. One way to make selected elements stand out is to place them in a central position. Linked elements must be aligned on the bottom based on the representation to make their presence evident. The remaining secondary details can be presented in a smaller size and with more subdued colours.

6.6 Discussion

To promote the quality of medical studies' results, researchers must rely on tools that help them in each decision-making process. The availability of biomedical database catalogues is an asset for searching biomedical data. In this section, some typical problems in using database catalogues are discussed, approaching the use of appropriate visualisations for semantic data.

6.6.1 Impact of Data Visualisations and Interactive Filtering

When researchers seek information to support a medical study, the interest is in knowing whether certain concepts are present in the available databases. In addition, they also need to know if the number of records is enough to support their work. For example, the Clinical Practice Research Datalink (CPRD) offers anonymised UK medical records, enabling the exploration of multiple dimensions such as demographics, symptoms and diagnoses (Herrett et al., 2015). The selection process is streamlined when viewing the percentages of records for each database concept at a glance is possible. Treemaps allow visualisation of how the records are distributed by different concepts, facilitating decision-making.

To set up a cohort, researchers need to understand the relationships between the concepts to iteratively improve the selection process and move forward with confidence in the choices they make. For instance, Huang et al. (2020) relates epidemiological, clinical, laboratory, and radiological data to study COVID-19 treatments and outcomes. Researchers are also interested in navigating between entities and exploring the ontology's connection network. They also hope to interact with the visualisation to fine-tune value windows or study if any concept is more closely linked to others. They can also use a layout that adds a value selection panel to this representation to fine-tune specific parameters. Ultimately, researchers can focus on a single node and explore in more detail the relationships it participates in for a given domain.

Timeline data often profoundly impact the quality of studies. For example, Esteban-Gil et al. (2017) consider the temporal dimension in using a semantic repository about cancer patients. The availability of historical data allows for refining the data of interest and noticing trends. The temporal dimension can be defined more or less blindly. When guided, visualisation facilitates the perception of hidden aspects, such as moments in time where data for a given dimension do not exist. Using a semantic data visualisation that allows exploring the time dimension should enable the choice of database versions that best suit the study's design. This way, the variability of updating different data sources is minimised, and the selection of data related to the concepts of interest is

improved.

The importance of comparative analysis of biomedical databases to conduct observational studies has been highlighted by several initiatives, such as the EMIF project or the Observational Health Data Sciences and Informatics (OHDSI) initiative (Hripcsak et al., 2015). OHDSI is an international, interdisciplinary, multi-stakeholder project to develop applications to access and analyse large-scale observational health data. The approaches described so far focus on exploring the metadata of a particular biomedical database about which researchers want to form an inclusion or exclusion opinion in a specific observational study. However, in a multicentre study, it is necessary to have a network view of the set of available databases to conduct a comparative analysis of the different options. Some authors have already contributed to this aim, like the Alzheimer’s disease community that created strategies to standardise distinct datasets and provide uniform methods to analyse them (Almeida et al., 2021).

Comparing one or more databases is central to performing multicentre studies, such as patient-level prediction studies. For example, Reps et al. (2021) use multiple health-care databases to reproduce two prediction models, one on type 2 diabetes and the other on dementia. A desirable way of deciding on data to train a model is, for instance, the possibility of comparatively studying the hierarchical structure of several databases. The availability of visual tools saves time and gives greater security in decision-making. In short, users have an advantage in seeing graphical representations in the form of a tree, as proposed before.

A researcher analysing the metadata of a given database may want to explore whether a particular concept is referred to in other databases. In this case, the idea is to focus on this single topic and search for it in other data sources. For instance, platforms for aggregating information on rare diseases usually collect data from distinct sources. The Diseasecard platform, which is one of these platforms, adopts graph representation and offers a navigation tree to examine networks of proteomic data and medical ontologies (Sequeira et al., 2021). With a graph-like representation, it is possible to select the node representing a concept for a particular database and navigate to find equivalents in another. The visual representation of nodes and relationships is the most intuitive way to explore new data from a previously defined concept for this type of navigation.

The selection of data based on a geographical criterion makes it possible to study specific populations. This is very common in multicentre studies, and sometimes the data is desired to belong to a particular country or set of countries. For instance, Morales et al. (2021) conducted a study using Spanish databases to identify renin-angiotensin system blockers and their susceptibility to COVID-19. However, making

choices using visual artefacts can be helpful, especially when keeping other seemingly less critical information visible. Access to a graphical representation of this type makes it possible to see several alternatives without losing focus on additional information that may be interesting to explore.

In medical studies, it is essential to have information on the types of databases per country and the number of records relating to a given concept, such as the number of patients with a given pathology. Researchers can quickly access this information using a heat map that crosses geographical data and the type of patients studied. Despite the multiple visualisation proposals discussed being a powerful tool to support researchers' decision-making in their search for biomedical data, several challenges remain.

6.6.2 Open Challenges and Future Directions

Semantic data visualisation is a subject that continues to raise different challenges depending on the volume of data, the complexity of ontologies, and the type of knowledge to be described. Dimensionality is critical for semantic networks with very high numbers of nodes and relationships, making visualisations hard to interpret. It is necessary to explore the creation of new algorithms based on the semantic network topology to reduce the weight of dimensionality in the representation of semantic data (Dadzie and Pietriga, 2016). In one of the EMIF Catalogue communities, the Alzheimer's disease community has a structure for collecting datasets' information composed of more than 430 concepts (Bos et al., 2018). In some views, with this number of concepts combined with a large number of registered datasets, performing a complete analysis with the traditional views is challenging for the researcher. However, the alternative, using a matrix view, is no better. Therefore, investing efforts in segmenting the information by adopting and implementing the visualisations described in Figures 6.13 and 6.14 would increase the system's usability.

New challenges arise when the use of multiple ontologies is required. Difficulties are compounded by annotation heterogeneity, which leads to the need to identify different terminologies for equal entities. In health database catalogues, this is common due to the existence of many domain-specific ontologies, such as the Human Phenotype Ontology (HPO) for phenotypic abnormalities and diseases (Köhler et al., 2016), the Gene Ontology (GO) for gene functions (Gene Ontology Consortium, 2016), and the Ontology for Biomedical Investigations (OBI) for scientific investigations (Brinkman et al., 2010), among others. There are reports of the use of service-oriented architectures to help in the efficient discovery of heterogeneous datasets in other domains (Zeshan et al., 2017). The most promising research directions on semantic similarity in the health domain point to the use of ontology embeddings in supervised learning ap-

proaches (Kulmanov et al., 2020).

Linking data over multiple semantic databases allows the creation of rich scenarios for questioning and visualising data. In this scenario, federated queries to obtain the desired information can be performed. However, performing federated queries remains challenging. This problem can be tackled by creating new indexing strategies and query processing schemes (Wylot et al., 2018).

A topic that sometimes goes unnoticed is privacy issues in publicly released catalogue data. There are already some algorithms to ensure data privacy for tabular data presentations (Sweeney, 2002; Machanavajjhala et al., 2007). However, this topic was not thoroughly studied when focusing on exposing the maximum knowledge from biomedical datasets using the proposed visualisations. Therefore, solid strategies are still needed to ensure the privacy aspects of semantic biomedical data, namely when the goal is to balance between maximum exposure, the client's goal, and minimal disclosure of information, the provider's concern.

The proposed time-evolving semantic data charts present advantages for researchers carrying out medical studies that depend on the careful selection of databases. However, there are some challenges in implementing and adopting the proposed visualisations. Concept evolution characterisation is challenging because it implies keeping a succession of states that allows the traceability of this evolution. It also means the need to compare different versions of the same ontology. One is faced with this scenario when evaluating data at the level of a single database. At the database level, it would be desirable for two versions of the ontology to see the operations of adding and deleting semantic entities. The visualisation of these two basic operations becomes complex when all the ontology elements overlap. Cardoso et al. (2020) solve this problem by building a historical knowledge graph that collects data related to all critical semantic operations: add, delete, split, move concepts, relationships or attributes. However, this problem still lacks an adequate solution, and its resolution would allow more informative visualisations.

6.7 Summary

The correct selection of databases to conduct a medical study may influence its success. Some studies could not be concluded due to a lack of a substantial number of subjects. However, researchers may simplify this process using adequate strategies to represent each database's characteristics.

Besides the more traditional querying and navigation techniques, interacting with semantic data using a set of visual artefacts was proposed, i.e., treemaps, graphs,

dendrograms, heat maps, and temporal charts. This kind of visualisation helps the exploration of database catalogues in greater depth, enabling analysis at multiple dimensions.

Data scarcity is a drawback when conducting observational medical studies. Considering historical data from the concept evolution of biomedical vocabularies can expand the range of data choices when using database catalogues. Information visualisation mechanisms are needed to facilitate decision-making, allowing for a more detailed view of the evolution of concepts in a database. It is also essential to compare different databases using the most appropriate data depictions.

The proposals were driven by the challenges of searching databases from a catalogue. The catalogue that guided the work contains metadata from biomedical databases with more than 1,000 fingerprints from 11 communities registered in the system. As a result, semantic data views at the database and network levels were proposed. This analysis pointed out future directions to develop new frameworks for representing semantic-based information. Although some of the proposed visualisations can be adopted using the available open-source solutions, the aim was to identify strategies that take the most significant advantage of the data using such visualisations. Some of the proposals may require implementing new features or components in such frameworks. Still, one of the issues that computational researchers have when trying to advance in the field of semantic data visualisation is the availability of practical and real use cases where new visualisations may have a significant impact.

Chapter 7

Conclusions and Future Work

Data integration and interoperability are problems that cause great concern in the scientific fields that depend heavily on the production and treatment of data. Nowadays, this reality is transversal to many scientific activities, namely life sciences. A commonly accepted way to alleviate these difficulties is using semantic data.

Creating large amounts of semantic data has led to multiple online repositories. However, the issue of creating and publishing these databases poses problems that need to be resolved for the benefit of users. On the other hand, the fact that standard users are not proficient in using formal query languages prevents them from effectively using these solutions.

This work aimed to solve the problem of creating and accessing semantic data by standard users, namely research domain specialists non-erudite in handling formal questioning languages.

7.1 Outcomes

A systematic literature review of the state-of-the-art KBQA systems was carried out, classifying the identified systems into four types of architecture. There are proposals that use classical semantic parsers to convert natural language questions into queries in a formal language, such as SPARQL. Another architectural type replaces part of this pipeline by direct subgraph lookup. Systems that use templates were also mentioned, and, finally, end-to-end systems that dispense entirely with the conversion of the question in natural language into a formal language. This work was published in the following paper:

Arnaldo Pereira, Alina Trifan, Rui Pedro Lopes, José Luís Oliveira, “Systematic review of question answering over knowledge bases”, *IET Software*, 2021, pp.1-13. <https://doi.org/10.1049/iet-soa.2020.0101>

[//doi.org/10.1049/sfw2.12028](https://doi.org/10.1049/sfw2.12028).

The problem of creating and publishing semantic data was approached considering the FAIR principles. On the one hand, one has to guarantee that the data can be found in search engines, which implies the creation of mechanisms that allow it to be seen by crawlers and indexed conveniently. The data must also have a suitable format that provides interoperability with other data available on the web. It is also necessary that data repositories are accessible under conditions established transparently and recoverable using open standards. These reflections were translated into the paper:

Arnaldo Pereira, Rui Pedro Lopes, José Luís Oliveira, “SCALEUS-FD: a FAIR data tool for biomedical applications”, *BioMed Research International*, 2020, pp.1-8. <https://doi.org/10.1155/2020/3041498>.

The result of this proposal was embedded in a tool for transforming and enriching semantic data and is freely available at:

<https://github.com/bioinformatics-ua/scaleus-fair>

A plugin for querying semantic data was implemented and applied to querying meta-data registered in a catalogue of biomedical databases. The code is available at:

<https://github.com/bioinformatics-ua/BioKBQA>

The ideas that support the implementation are developed in the paper:

Arnaldo Pereira, João Rafael Almeida, Rui Pedro Lopes, José Luís Oliveira, “Querying semantic catalogues of biomedical databases.” (Submitted.)

The application of the KBQA solution to data from patients with Huntington’s disease was reported in the conference paper:

Arnaldo Pereira, Rui Pedro Lopes, José Luís Oliveira, “Easing the questioning of semantic biomedical data”, In: *IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, 2021, pp.384-388. <https://doi.org/10.1109/CBMS52027.2021.00044>.

Several proposals were made for visualising semantic data to support decision-making in choosing biomedical databases, using a catalogue. These visualisations were discussed in the following two papers:

Arnaldo Pereira, João Rafael Almeida, Rui Pedro Lopes, José Luís Oliveira, “Visualising time-evolving semantic biomedical data”, In: *IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, 2022, pp.264-269. <https://doi.org/10.1109/CBMS52027.2022.00044>.

[//doi.org/10.1109/CBMS55023.2022.00053](https://doi.org/10.1109/CBMS55023.2022.00053).

Arnaldo Pereira, João Rafael Almeida, Rui Pedro Lopes, José Luís Oliveira, “Semantic data visualisation for biomedical database catalogues.” (Submitted.)

7.2 Future Work

Several factors affect the quality of semantic data querying systems that have not been explored in the scope of this work and are lines of future work that deserve close attention.

- A limitation of the proposal is that it does not deal well with semantic database incompleteness. One first line of investigation concerns data quality and its impact on the performance of KBQA systems. Semantic data can be generated in several ways, namely by automated mining of entities and relationships from text. This process can cause corrupted or incomplete data, and it is necessary to alleviate these difficulties. The use of hybrid systems that complement incomplete information using text information has been explored. Still, this problem remains open, and its solution could significantly impact the quality of KBQA systems.
- Another limitation related to query systems in general that also impact the usability of KBQA is the ability to guide the user in the formulation of the question. A possible solution for this topic of great interest that is still open can be constructing conversational systems.

Appendix A

Systematic Review Publications

Table A.1: List of publications included in the systematic review of KBQA.

ID	Paper
1	Answering questions with complex semantic constraints on open knowledge bases (Yin et al., 2015)
2	Applying semantic parsing to question answering over linked data: addressing the lexical gap (Hakimov et al., 2015)
3	HAWK - hybrid question answering using linked data (Usbeck et al., 2015)
4	How to build templates for RDF question/answering - An uncertain graph similarity join approach (Zheng et al., 2015)
5	ISOFT at QALD-5: Hybrid question answering system over linked data and text data (Park et al., 2015)
6	More accurate question answering on Freebase (Bast and Haussmann, 2015)
7	QAnswer - Enhanced entity matching for question answering over linked data (Ruseti et al., 2015)
8	Question answering over Freebase with multi-column convolutional neural networks (Dong et al., 2015)
9	Question answering via phrasal semantic parsing (Xu et al., 2014)
10	Semantic parsing via staged query graph generation: question answering with knowledge base (Yih et al., 2015)
11	SemGraphQA@QALD-5: LIMSIS participation at QALD-5@CLEF (Beaumont et al., 2015)
12	SINA: Semantic interpretation of user queries for question answering on interlinked data (Shekarpour et al., 2015)

(continued on next page)

Table A.1 (*continued*)

ID	Paper
13	TR Discover: A natural language interface for querying and analysing interlinked datasets (Song et al., 2015)
14	AskNow: A framework for natural language query formalization in SPARQL (Dubey et al., 2016)
15	CFO: Conditional focussed neural question answering with large-scale knowledge bases (Dai et al., 2016)
16	Character-level question answering with attention (He and Golub, 2016)
17	Constraint-based question answering with knowledge graph (Bao et al., 2016)
18	GRU-RNN based question answering over knowledge base (Chen et al., 2016)
19	Hybrid question answering over knowledge base and free text (Xu et al., 2016a)
20	Knowledge base question answering based on deep learning models (Xie et al., 2016)
21	Neural generative question answering (Yin et al., 2016)
22	Qanary - A methodology for vocabulary-driven open question answering systems (Both et al., 2016)
23	QuerioDALI: Question answering over dynamic and linked knowledge graphs (Lopez et al., 2016)
24	Question answering on Freebase via relation extraction and textual evidence (Xu et al., 2016b)
25	The value of semantic parse labelling for knowledge base question answering (Yih et al., 2016)
26	When a knowledge base is not enough: Question answering over knowledge bases with external text data (Savenkov and Agichtein, 2016)
27	An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge (Hao et al., 2017)
28	Automated template generation for question answering over knowledge graphs (Abujabal et al., 2017)
29	End-to-end representation learning for question answering with weak supervision (Sorokin and Gurevych, 2017)
30	Improved neural relation detection for knowledge base question answering (Yu et al., 2017)

(continued on next page)

Table A.1 (*continued*)

ID	Paper
31	Introducing feedback in Qanary: How users can Interact with QA systems (Diefenbach et al., 2017b)
32	KBQA: Learning question answering over QA corpora and knowledge bases (Cui et al., 2017)
33	Matching natural language relations to knowledge graph properties for question answering (Mulang et al., 2017)
34	Natural language supported relation matching for question answering with knowledge graphs (Li et al., 2017)
35	Neural network-based question answering over knowledge graphs on word and character levels (Lukovnikov et al., 2017)
36	QAESTRO—semantic-based composition of question answering pipelines (Singh et al., 2017)
37	Querying biomedical linked data with natural language questions (Hamon et al., 2017)
38	Question answering on knowledge bases and text using universal schema and memory networks (Das et al., 2017)
39	Trill: A reusable front-end for QA systems (Diefenbach et al., 2017a)
40	An attention-based word-level interaction model for knowledge base relation detection (Zhang et al., 2018)
41	Answering natural language questions by subgraph matching over knowledge graphs (Hu et al., 2018a)
42	Formal query generation for question answering over knowledge bases (Zafar et al., 2018)
43	Frankenstein: A platform enabling reuse of question answering components (Singh et al., 2018a)
44	Never-ending learning for open-domain question answering over knowledge bases (Abujabal et al., 2018)
45	Novel knowledge-based system with relation detection and textual evidence for question answering research (Zheng et al., 2018a)
46	Question answering over knowledge graphs: Question understanding via template decomposition (Zheng et al., 2018b)
47	Svega: Answering natural language questions over knowledge base with semantic matching (Li et al., 2018)

(continued on next page)

Table A.1 (*continued*)

ID	Paper
48	Why reinvent the wheel: Let’s build question answering systems together (Singh et al., 2018b)
49	Answer-enhanced path-aware relation detection over knowledge base (Chen et al., 2019)
50	Complex query augmentation for question answering over knowledge graphs (Abdelkawi et al., 2019)
51	ComQA: Question answering over knowledge base via semantic matching (Jin et al., 2019)
52	Deep query ranking for question answering over knowledge bases (Zafar et al., 2019)
53	Handling modifiers in question answering over knowledge graphs (Siciliani et al., 2019)
54	Knowledge base question answering with a matching-aggregation model and question-specific contextual relations (Lan et al., 2019)
55	Knowledge base question answering with attentive pooling for question representation (Wang et al., 2019)
56	Learning to answer complex questions over knowledge bases with query composition (Bhutani et al., 2019)
57	Learning to rank query graphs for complex question answering over knowledge graphs (Maheshwari et al., 2019)
58	Message passing for complex question answering over knowledge graphs (Vakulenko et al., 2019)
59	Pretrained transformers for simple question answering over knowledge graphs (Lukovnikov et al., 2019)
60	A BERT-based approach with relation-aware attention for knowledge base question answering (Luo et al., 2020a)
61	A state-transition framework to answer complex questions over knowledge base (Hu et al., 2018b)
62	Data-driven construction of SPARQL queries by approximate question graph alignment in question answering over knowledge graphs (Bakhshi et al., 2020)
63	Exploring sequence-to-sequence models for SPARQL pattern composition (Panchbhai et al., 2020)

(continued on next page)

Table A.1 (*continued*)

ID	Paper
64	Formal query building with query structure prediction for complex question answering over knowledge base (Chen et al., 2020)
65	Improving question answering over incomplete KBs with knowledge-aware reader (Xiong et al., 2019)
66	Knowledge base question answering via encoding of complex query graphs (Luo et al., 2020b)

Appendix B

KBQA Benchmark Datasets Data Samples

Table B.1: Data samples from QALD, LC-QuAD, Free917, WebQuestions, WebQuestionsSP, SimpleQuestions, ComQA, and BioASQ benchmark datasets.

	Question, utterance
Benchmark	----- Target, logic form, answer
QALD	"language" : "en", "string" : "List all boardgames by GMT.", ----- "sparql" : "PREFIX dbo: <http://dbpedia.org/ontology/> (...) SELECT ?uri WHERE { ?uri dbo:publisher res:GMT_Games }"
LC-QuAD	"question": "What periodical literature does Delta Air Lines use as a moutpiece?", (...) "paraphrased_question": "What is Delta Air Line's periodical literature mouthpiece?" ----- "sparql_wikidata": "select distinct ?obj where { wd:Q188920 wdt:P2813 ?obj . ?obj wdt:P31 wd:Q1002697 }", "sparql_dbpedia18": "select distinct ?obj where (...)

(continued on next page)

Table B.1 (*continued*)

	Question, utterance
Benchmark	-----
	Target, logic form, answer
Free917	<pre> {"utterance": "what fuel does an internal combustion engine use", "targetFormula": "(!fb:engineering.engine.energy_source fb:en.internal_combustion_engine)"}, </pre>
WebQ	<pre> "utterance": "what is the name of justin bieber brother?" "targetValue": "(list (description \"Jazmyn Bieber\") (description \"Jaxon Bieber\"))" </pre>
WebQSP	<pre> "RawQuestion": "what is the name of justin bieber brother?", "ProcessedQuestion": "what is the name of justin bieber brother", "SPARQL": "PREFIX ns: <http://rdf.freebase.com/ns/>\nSELECT DISTINCT ?x\nWHERE \nFILTER (?x != ns:m.06w2sn5)\nFILTER (!isLiteral(?x) OR lang(?x) = \" OR langMatches(lang(?x), 'en'))\nns:m.06w2sn5 ns:people.person.sibling_s ?y .\n?y ns:people.sibling_relationship.sibling ?x .\n?x ns:people.person.gender ns:m.05zppz .\n\n", </pre>
SimpleQ	<pre> www.freebase.com/fictional_universe/ fictional_character/character_created_by www.freebase.com/m/037w1 what American cartoonist is the creator of andy lippincott </pre>
	----- www.freebase.com/m/05kg30

(continued on next page)

Table B.1 (*continued*)

	Question, utterance
Benchmark	-----
	Target, logic form, answer
ComQA	<pre>"questions": ["who plays james potter in the harry potter films?", "who is james potter the harry potter father?"], ----- "answers": ["https://en.wikipedia.org/wiki/robbie_jarvis"]</pre>
BioASQ	<pre>"body": "Which 2 medications are included in the Qsymia pill?", ----- { "p": "http://www.w3.org/2008/05/skos-xl#altLabel", "s": "http://linkedlifedata.com/resource/umls/id/ C0013227", "o": "http://linkedlifedata.com/resource/umls/label/ A18591068" }, (...)</pre>

References

- Abad-Navarro, F., C. Martínez-Costa, and J. T. Fernández-Breis (2021). “Semanky: a semantics-driven approach for querying RDF repositories using keywords.” In: *IEEE Access* 9, pp. 91282–91302. DOI: 10.1109/ACCESS.2021.3091413.
- Abdelkawi, A., H. Zafar, M. Maleshkova, and J. Lehmann (2019). “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)*, pp. 571–587. DOI: 10.1007/978-3-030-33246-4_36.
- Abujabal, A., R. Saha Roy, M. Yahya, and G. Weikum (2018). “Never-ending learning for open-domain question answering over knowledge bases.” In: *Proceedings of the 27th World Wide Web Conference (WWW)*, pp. 1053–1062. DOI: 10.1145/3178876.3186004.
- Abujabal, A., M. Yahya, M. Riedewald, and G. Weikum (2017). “Automated template generation for question answering over knowledge graphs.” In: *Proceedings of the 26th International Conference on World Wide Web (WWW)*, pp. 1191–1200. DOI: 10.1145/3038912.3052583.
- Affolter, K., K. Stockinger, and A. Bernstein (2019). “A comparative survey of recent natural language interfaces for databases.” In: *The VLDB Journal* 28.5, pp. 793–819. DOI: 10.1007/s00778-019-00567-8.
- Agnihotri, M. and A. Chug (2021). “Analyzing the relationship between software metrics and bad smells using critical metric value (CMV).” In: *Proceedings of the 13th International Conference on Contemporary Computing (IC3)*, pp. 450–456. DOI: 10.1145/3474124.3474193.

- Aigner, W., S. Miksch, H. Schumann, and C. Tominski (2011). “Survey of visualization techniques.” In: *Visualization of Time-Oriented Data*. Human-Computer Interaction Series. London: Springer. Chap. 7, pp. 147–254. DOI: 10.1007/978-0-85729-079-3_7.
- Almeida, J. R., O. Fajarda, A. Pereira, and J. L. Oliveira (2019). “Strategies to access patient clinical data from distributed databases.” In: *Proceedings of the 12th International Conference on Health Informatics (HEALTHINF)*, pp. 466–473. DOI: 10.5220/0007576104660473.
- Almeida, J. R., E. Monteiro, L. B. Silva, A. P. Sierra, and J. L. Oliveira (2020). “A recommender system to help discovering cohorts in rare diseases.” In: *Proceedings of the 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 25–28. DOI: 10.1109/CBMS49503.2020.00012.
- Almeida, J. R., R. Ribeiro, and J. L. Oliveira (2018). “A modular workflow management framework.” In: *Proceedings of the 11th International Conference on Health Informatics (HealthInf)*, pp. 414–421. DOI: 10.5220/0006583104140421.
- Almeida, J. R., L. B. Silva, I. Bos, P. J. Visser, and J. L. Oliveira (2021). “A methodology for cohort harmonisation in multicentre clinical research.” In: *Informatics in Medicine Unlocked* 27, pp. 1–9. DOI: 10.1016/j.imu.2021.100760.
- Androutsopoulos, I., G. D. Ritchie, and P. Thanisch (1995). “Natural language interfaces to databases – an introduction.” In: *Natural Language Engineering* 1.1, pp. 29–81. DOI: 10.1017/S135132490000005X.
- Angles, R., M. Arenas, P. Barceló, A. Hogan, J. Reutter, and D. Vrgoč (2017). “Foundations of modern query languages for graph databases.” In: *ACM Computing Surveys* 50.5, pp. 1–40. DOI: 10.1145/3104031.
- Angles, R. and C. Gutierrez (2008). “Survey of graph database models.” In: *ACM Computing Surveys* 40.1, pp. 1–39. DOI: 10.1145/1322432.1322433.
- Angles, R., H. Thakkar, and D. Tomaszuk (2020). “Mapping RDF databases to property graph databases.” In: *IEEE Access* 8, pp. 86091–86110. DOI:

- 10.1109/ACCESS.2020.2993117.
- Asiaee, A. H., T. Minning, P. Doshi, and R. L. Tarleton (2015). “A framework for ontology-based question answering with application to parasite immunology.” In: *Journal of Biomedical Semantics* 6.1, pp. 1–25. DOI: 10.1186/s13326-015-0029-x.
- Azad, H. K., A. Deepak, and A. Azad (2021). “LOD search engine: a semantic search over linked data.” In: *Journal of Intelligent Information Systems*, pp. 1–21. DOI: 10.1007/s10844-021-00687-0.
- Bach, B., C. Shi, N. Heulot, T. Madhyastha, T. Grabowski, and P. Dragicevic (2016). “Time curves: folding time to visualize patterns of temporal evolution in data.” In: *IEEE Transactions on Visualization and Computer Graphics* 22.1, pp. 559–568. DOI: 10.1109/TVCG.2015.2467851.
- Bakhshi, M., M. Nematbakhsh, M. Mohsenzadeh, and A. M. Rahmani (2020). “Data-driven construction of SPARQL queries by approximate question graph alignment in question answering over knowledge graphs.” In: *Expert Systems with Applications* 146, pp. 1–19. DOI: 10.1016/j.eswa.2020.113205.
- Bao, J., N. Duan, Z. Yan, M. Zhou, and T. Zhao (2016). “Constraint-based question answering with knowledge graph.” In: *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*. URL: <https://aclanthology.org/C16-1236>, pp. 2503–2514.
- Bast, H. and E. Haussmann (2015). “More accurate question answering on Freebase.” In: *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pp. 1431–1440. DOI: 10.1145/2806416.2806472.
- Beaumont, R., B. Grau, and A.-L. Ligozat (2015). “SemGraphQA@QALD-5: LIMSI participation at QALD-5@CLEF.” In: *Proceedings of the 16th Conference and Labs of the Evaluation Forum (CLEF)*. URL: <http://ceur-ws.org/Vol-1391/164-CR.pdf>, pp. 1–10.
- Beckett, D., T. Berners-Lee, E. Prud’hommeaux, and G. Carothers (2014). *RDF 1.1 Turtle: Terse RDF Triple Language. W3C recommendation*. URL: <https://www.w3.org/TR/turtle/>

[//www.w3.org/TR/turtle/](http://www.w3.org/TR/turtle/).

- Berant, J., A. Chou, R. Frostig, and P. Liang (2013). “Semantic parsing on Freebase from question-answer pairs.” In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. URL: <https://aclanthology.org/D13-1160>, pp. 1533–1544.
- Berners-Lee, T. (2006). *Linked Data*. URL: <https://www.w3.org/DesignIssues/LinkedData.html>.
- Berners-Lee, T., J. Hendler, and O. Lassila (2001). “The Semantic Web.” In: *Scientific American* 284.5, pp. 34–43. DOI: 10.1038/scientificamerican0501-34.
- Bhutani, N., X. Zheng, and H. V. Jagadish (2019). “Learning to answer complex questions over knowledge bases with query composition.” In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 739–748. DOI: 10.1145/3357384.3358033.
- Böhlen, M. H., A. Dignös, J. Gamper, and C. S. Jensen (2017). “Temporal data management - an overview.” In: *Proceedings of the 7th European Summer School on Business Intelligence and Big Data (eBISS)*, pp. 51–83. DOI: 10.1007/978-3-319-96655-7_3.
- Bollacker, K., C. Evans, P. Paritosh, T. Sturge, and J. Taylor (2008). “Freebase: a collaboratively created graph database for structuring human knowledge.” In: *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pp. 1247–1250. DOI: 10.1145/1376616.1376746.
- Bonino da Silva Santos, L. O., M. Wilkinson, A. Kuzniar, R. Kaliyaperumal, M. Thompson, M. Dumontier, and K. Burger (2016). “FAIR data points supporting big data interoperability.” In: *Enterprise Interoperability in the Digitized and Networked Factory of the Future*. URL: <https://tinyurl.com/ypt7nspf>. ISTE Press, pp. 270–279.
- Bordes, A., N. Usunier, S. Chopra, and J. Weston (2015). “Large-scale simple question answering with memory networks.” In: *CoRR* abs/1506.02075, pp. 1–10. DOI:

10.48550/arXiv.1506.02075.

- Borst, W. N. (1997). “Construction of engineering ontologies for knowledge sharing and reuse.” URL: <http://doc.utwente.nl/17864/>. PhD thesis. University of Twente.
- Bos, I., S. Vos, R. Vandenberghe, P. Scheltens, S. Engelborghs, G. Frisoni, J. L. Molinuevo, A. Wallin, A. Lleó, J. Popp, P. Martinez-Lage, A. Baird, R. Dobson, C. Legido-Quigley, K. Slegers, C. V. Broeckhoven, L. Bertram, M. t. Kate, F. Barkhof, H. Zetterberg, S. Lovestone, J. Streffer, and P. J. Visser (2018). “The EMIF-AD Multimodal Biomarker Discovery study: design, methods and cohort characteristics.” In: *Alzheimer’s Research & Therapy* 10.1, pp. 1–9. DOI: 10.1186/s13195-018-0396-5.
- Both, A., D. Diefenbach, K. Singh, S. Shekarpour, D. Cherix, and C. Lange (2016). “Qanary - a methodology for vocabulary-driven open question answering systems.” In: *Proceedings of the 13th European Semantic Web Conference (ESWC)*, pp. 625–641. DOI: 10.1007/978-3-319-34129-3_38.
- Brickley, D., M. Burgess, and N. Noy (2019). “Google Dataset Search: building a search engine for datasets in an open web ecosystem.” In: *Proceedings of the 28th World Wide Web Conference (WWW)*, pp. 1365–1375. DOI: 10.1145/3308558.3313685.
- Brickley, D. and R. V. Guha (2014). *RDF Schema 1.1. W3C recommendation*. URL: <http://www.w3.org/TR/rdf-schema/>.
- Brinkman, R., M. Courtot, D. Derom, J. Fostel, Y. He, P. Lord, J. Malone, H. Parkinson, B. Peters, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, L. Soldatova, C. Stoeckert, J. Turner, and J. Zheng (2010). “Modeling biomedical experimental processes with OBI.” In: *Journal of biomedical semantics* 1, pp. 1–11. DOI: 10.1186/2041-1480-1-S1-S7.
- Cai, Q. and A. Yates (2013). “Large-scale semantic parsing via schema matching and lexicon extension.” In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. URL: <https://aclanthology.org/P13-1042>, pp. 423–433.

- Callahan, A., J. Cruz-Toledo, P. Ansell, and M. Dumontier (2013). “Bio2RDF Release 2: improved coverage, interoperability and provenance of life science linked data.” In: *Proceedings of the 10th International Conference on The Semantic Web: Semantics and Big Data (ESWC)*, pp. 200–212. DOI: 10.1007/978-3-642-38288-8_14.
- Cardoso, S. D., M. Silveira, and C. Pruski (2020). “Construction and exploitation of an historical knowledge graph to deal with the evolution of ontologies.” In: *Knowledge-Based Systems* 194, p. 105508. DOI: 10.1016/j.knosys.2020.105508.
- Catarci, T., M. F. Costabile, S. Levialdi, and C. Batini (1997). “Visual query systems for databases: a survey.” In: *Journal of Visual Languages & Computing* 8.2, pp. 215–260. DOI: 10.1006/jv1c.1997.0037.
- Chen, B., Y. Ding, and D. J. Wild (2012). “Assessing drug target association using semantic linked data.” In: *PLOS Computational Biology* 8.7, pp. 1–10. DOI: 10.1371/journal.pcbi.1002574.
- Chen, B., X. Dong, D. Jiao, H. Wang, Q. Zhu, Y. Ding, and D. Wild (2010). “Chem2Bio2RDF: A semantic framework for linking and data mining chemogenomic and systems chemical biology data.” In: *BMC Bioinformatics* 11.1, pp. 1–13. DOI: 10.1186/1471-2105-11-255.
- Chen, D., M. Yang, H.-T. Zheng, Y. Li, and Y. Shen (2019). “Answer-enhanced path-aware relation detection over knowledge base.” In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1021–1024. DOI: 10.1145/3331184.3331328.
- Chen, S., J. Wen, and R. Zhang (2016). “GRU-RNN based question answering over knowledge base.” In: *Proceedings of the 1st China Conference on Knowledge Graph and Semantic Computing (CCKS)*, pp. 80–91. DOI: 10.1007/978-981-10-3168-7_8.
- Chen, Y., H. Li, Y. Hua, and G. Qi (2020). “Formal query building with query structure prediction for complex question answering over knowledge base.” In: *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3751–3758. DOI: 10.24963/ijcai.2020/519.

- Cheng, H. G. and M. R. Phillips (2014). “Secondary analysis of existing data: opportunities and implementation.” In: *Shanghai Archives of Psychiatry* 26.6, pp. 371–375. DOI: 10.11919/j.issn.1002-0829.214171.
- Chiueh, T.-c. and D. Pilania (2005). “Design, implementation, and evaluation of a repairable database management system.” In: *Proceedings of the 21st International Conference on Data Engineering (ICDE)*, pp. 1024–1035. DOI: 10.1109/ICDE.2005.49.
- Cockburn, A., A. Karlson, and B. B. Bederson (2009). “A review of overview+detail, zooming, and focus+context interfaces.” In: *ACM Computing Surveys* 41.1, pp. 1–31. DOI: 10.1145/1456650.1456652.
- Cui, W., Y. Xiao, H. Wang, Y. Song, S.-w. Hwang, and W. Wang (2017). “KBQA: learning question answering over QA corpora and knowledge bases.” In: *Proceedings of the 43rd International Conference on Very Large Data Bases (VLDB)* 10.5, pp. 565–576. DOI: 10.14778/3055540.3055549.
- Cyganiak, R., D. Wood, and M. Lanthaler (2014). *RDF 1.1 concepts and abstract syntax. W3C recommendation*. URL: <https://www.w3.org/TR/rdf11-concepts/>.
- Dadzie, A.-S. and E. Pietriga (2016). “Visualisation of Linked Data - reprise.” In: *Semantic Web* 8.1, pp. 1–21. DOI: 10.3233/SW-160249.
- Dai, Z., L. Li, and W. Xu (2016). “CFO: conditional focused neural question answering with large-scale knowledge bases.” In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 800–810. DOI: 10.18653/v1/P16-1076.
- Daiber, J., M. Jakob, C. Hokamp, and P. N. Mendes (2013). “Improving efficiency and accuracy in multilingual entity extraction.” In: *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*, pp. 121–124. DOI: 10.1145/2506182.2506198.
- Das, R., M. Zaheer, S. Reddy, and A. McCallum (2017). “Question answering on knowledge bases and text using universal schema and memory networks.” In: *Proceedings of the 55th Annual Meeting of the Association for Computational*

- Linguistics (Volume 2: Short Papers)*, pp. 358–365. DOI: 10.18653/v1/P17-2057.
- Das, S., J. Srinivasan, M. Perry, E. I. Chong, and J. Banerjee (2014). “A tale of two graphs: property graphs as RDF in Oracle.” In: *Proceedings of 17th International Conference on Extending Database Technology (EDBT)*, pp. 762–773. DOI: 10.5441/002/edbt.2014.82.
- De Moor, G., 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, and P. Coorevits (2015). “Using electronic health records for clinical research: the case of the EHR4CR project.” In: *Journal of Biomedical Informatics* 53, pp. 162–173. DOI: 10.1016/j.jbi.2014.10.006.
- Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019). “BERT: pre-training of deep bidirectional transformers for language understanding.” In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.
- Diefenbach, D., S. Amjad, A. Both, K. Singh, and P. Maret (2017a). “Trill: a reusable front-end for QA systems.” In: *Proceedings of the 14th European Semantic Web Conference (ESWC)*, pp. 48–53. DOI: 10.1007/978-3-319-70407-4_10.
- Diefenbach, D., N. Hormozi, S. Amjad, and A. Both (2017b). “Introducing feedback in Qanary: how users can interact with QA systems.” In: *Proceedings of the 14th European Semantic Web Conference (ESWC)*, pp. 81–86. DOI: 10.1007/978-3-319-70407-4_16.
- Dimitrakis, E., K. Sgontzos, and Y. Tzitzikas (2020). “A survey on question answering systems over linked data and documents.” In: *Journal of Intelligent Information Systems* 55.2, pp. 233–259. DOI: 10.1007/s10844-019-00584-7.
- Djokic-Petrovic, M., V. Cvjetkovic, J. Yang, M. Zivanovic, and D. Wild (2017). “PIBAS FedSPARQL: a web-based platform for integration and exploration of bioinformatics datasets.” In: *Journal of Biomedical Semantics* 8, pp. 1–20. DOI: 10.1186/s13326-017-0151-z.

- Dong, L., F. Wei, M. Zhou, and K. Xu (2015). “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)*, pp. 260–269. DOI: 10.3115/v1/P15-1026.
- Drysdale, R., C. E. Cook, R. Petryszak, V. Baillie-Gerritsen, M. Barlow, 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, and J. McEntyre (2020). “The ELIXIR Core Data Resources: fundamental infrastructure for the life sciences.” In: *Bioinformatics* 36.8, pp. 2636–2642. DOI: 10.1093/bioinformatics/btz959.
- Dubey, M., D. Banerjee, A. Abdelkawi, and J. Lehmann (2019). “LC-QuAD 2.0: a large dataset for complex question answering over Wikidata and DBpedia.” In: *Proceedings of the International Semantic Web Conference (ISWC)*, pp. 69–78. DOI: 10.1007/978-3-030-30796-7_5.
- Dubey, M., S. Dasgupta, A. Sharma, K. Höffner, and J. Lehmann (2016). “AskNow: a framework for natural language query formalization in SPARQL.” In: *Proceedings of the 13th European Semantic Web Conference (ESWC)*, pp. 300–316. DOI: 10.1007/978-3-319-34129-3_19.
- Dürst, M. and M. Suignard (2005). *Internationalized Resource Identifiers (IRIs)*. URL: <http://www.ietf.org/rfc/rfc3987.txt>.
- Ehrlinger, L. and W. Wöß (2016). “Towards a definition of knowledge graphs.” In: *Proceedings of the 12th International Conference on Semantic Systems (SEMANTiCS)*. URL: <http://ceur-ws.org/Vol-1695/paper4.pdf>, pp. 1–4.
- Elbedweihy, K., S. Wrigley, P. Clough, and F. Ciravegna (2015). “An overview of semantic search evaluation initiatives.” In: *Journal of Web Semantics* 30.C, pp. 82–105. DOI: 10.1016/j.websem.2014.10.001.
- Elmqvist, N. and J. S. Yi (2015). “Patterns for visualization evaluation.” In: *Information Visualization* 14.3, pp. 250–269. DOI: 10.1177/1473871613513228.

- Esteban-Gil, A., J. Fernandez-Breis, and M. Boeker (2017). “Analysis and visualization of disease courses in a semantic enabled cancer registry.” In: *Journal of Biomedical Semantics* 8, pp. 1–16. DOI: 10.1186/s13326-017-0154-9.
- Fajarda, O., L. B. Silva, P. R. Rijnbeek, M. Van Speybroeck, and J. L. Oliveira (2018). “A methodology to perform semi-automatic distributed EHR database queries.” In: *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (HEALTHINF)*, pp. 127–134. DOI: 10.5220/0006579701270134.
- Fan, J., A. Kalyanpur, D. C. Gondek, and D. A. Ferrucci (2012). “Automatic knowledge extraction from documents.” In: *IBM Journal of Research and Development* 56.3, pp. 1–10. DOI: 10.1147/JRD.2012.2186519.
- Fernández, M., A. Gómez-Pérez, and N. Juristo (1997). “METHONTOLOGY: from ontological art towards ontological engineering.” In: *Proceedings of the Ontological Engineering AAAI-97 Spring Symposium Series*. URL: <https://www.aaai.org/Papers/Symposia/Spring/1997/SS-97-06/SS97-06-005.pdf>.
- Ferré, S. (2017). “Sparklis: an expressive query builder for SPARQL endpoints with guidance in natural language.” In: *Semantic Web* 8.3, pp. 405–418. DOI: 10.3233/SW-150208.
- Francis, L. P. and J. G. Francis (2017). “Data reuse and the problem of group identity.” In: *Studies in Law Politics and Society* 73, pp. 141–164. DOI: 10.1108/S1059-433720170000073004.
- Francis, N., A. Green, P. Guagliardo, L. Libkin, T. Lindaaker, V. Marsault, S. Planktikow, M. Rydberg, P. Selmer, and A. Taylor (2018). “Cypher: an evolving query language for property graphs.” In: *Proceedings of the 2018 International Conference on Management of Data*, pp. 1433–1445. DOI: 10.1145/3183713.3190657.
- Gall, C. S., S. Lukins, L. Etzkorn, S. Gholston, P. Farrington, D. Utley, J. Fortune, and S. Virani (2008). “Semantic software metrics computed from natural language design specifications.” In: *IET Software* 2.1, pp. 17–26. DOI: 10.1049/iet-sen:20070109.

- Garcia-Silva, A., J. M. Gomez-Perez, R. Palma, M. Krystek, S. Mantovani, F. Foglini, V. Grande, F. De Leo, S. Salvi, E. Trasatti, V. Romaniello, M. Albani, C. Silvagni, R. Leone, F. Marelli, S. Albani, M. Lazzarini, H. J. Napier, H. M. Glaves, T. Aldridge, C. Meertens, F. Boler, H. W. Loesch, C. Laney, M. A. Genazzio, D. Crawl, and I. Altintas (2019). “Enabling FAIR research in Earth Science through research objects.” In: *Future Generation Computer Systems* 98, pp. 550–564. DOI: 10.1016/j.future.2019.03.046.
- Gayo, J. E. L., D. F. Álvarez, and H. García-González (2018). “RDFShape: an RDF playground based on shapes.” In: *Proceedings of the 17th International Semantic Web Conference (ISWC)*. URL: <http://ceur-ws.org/Vol-2180/paper-35.pdf>, pp. 1–4.
- Gene Ontology Consortium (2016). “Expansion of the Gene Ontology knowledge-base and resources.” In: *Nucleic Acids Research* 45.D1, pp. D331–D338. DOI: 10.1093/nar/gkw1108.
- Goel, S., A. Broder, E. Gabrilovich, and B. Pang (2010). “Anatomy of the long tail: ordinary people with extraordinary tastes.” In: *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pp. 201–210. DOI: 10.1145/1718487.1718513.
- Golshan, B., A. Halevy, G. Mihaila, and W.-C. Tan (2017). “Data integration: after the teenage years.” In: *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pp. 101–106. DOI: 10.1145/3034786.3056124.
- Goodman, A., A. Pepe, A. W. Blocker, C. L. Borgman, K. Cranmer, M. Crosas, R. Di Stefano, Y. Gil, P. Groth, M. Hedstrom, D. W. Hogg, V. Kashyap, A. Mahabal, A. Siemiginowska, and A. Slavkovic (2014). “Ten simple rules for the care and feeding of scientific data.” In: *PLOS Computational Biology* 10.4, pp. 1–5. DOI: 10.1371/journal.pcbi.1003542.
- Green, B. F., A. K. Wolf, C. Chomsky, and K. Laughery (1961). “Baseball: an automatic question-answerer.” In: *Papers Presented at the May 9-11, 1961, Western Joint IRE-AIEE-ACM Computer Conference*, pp. 219–224. DOI:

10.1145/1460690.1460714.

- Groth, P., A. Loizou, A. J. G. Gray, C. Goble, L. Harland, and S. Pettifer (2014). “API-centric Linked Data integration: the Open PHACTS Discovery Platform case study.” In: *Journal of Web Semantics* 29, pp. 12–18. DOI: 10.2139/ssrn.3199140.
- Gruber, T. R. (1993). “A translation approach to portable ontology specifications.” In: *Knowledge Acquisition* 5.2, pp. 199–220. DOI: 10.1006/knac.1993.1008.
- Gubichev, A., S. Bedathur, S. Seufert, and G. Weikum (2010). “Fast and accurate estimation of shortest paths in large graphs.” In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 499–508. DOI: 10.1145/1871437.1871503.
- Guha, R., R. McCool, and E. Miller (2003). “Semantic search.” In: *Proceedings of the 12th International Conference on World Wide Web (WWW)*, pp. 700–709. DOI: 10.1145/775152.775250.
- Hakimov, S., C. Unger, S. Walter, and P. Cimiano (2015). “Applying semantic parsing to question answering over linked data: addressing the lexical gap.” In: *Proceedings of the 20th International Conference on Applications of Natural Language to Information Systems (NLDB)*, pp. 103–109. DOI: 10.1007/978-3-319-19581-0_8.
- Hamon, T., N. Grabar, and F. Mougín (2017). “Querying biomedical Linked Data with natural language questions.” In: *Semantic Web* 8.4, pp. 581–599. DOI: 10.3233/SW-160244.
- Hanspers, K., M. Kutmon, S. L. Coort, D. Digles, L. J. Dupuis, F. Ehrhart, F. Hu, E. N. Lopes, M. Martens, N. Pham, W. Shin, D. N. Slenter, A. Waagmeester, E. L. Willighagen, L. A. Winckers, C. T. Evelo, and A. R. Pico (2021). “Ten simple rules for creating reusable pathway models for computational analysis and visualization.” In: *PLOS Computational Biology* 17.8, pp. 1–14. DOI: 10.1371/journal.pcbi.1009226.
- Hao, Y., Y. Zhang, K. Liu, S. He, Z. Liu, H. Wu, and J. Zhao (2017). “An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge.” In: *Proceedings of the 55th Annual Meeting of the Association*

-
- for *Computational Linguistics (Volume 1: Long Papers)*, pp. 221–231. DOI: 10.18653/v1/P17-1021.
- Harris, S. and A. Seaborne (2013). *SPARQL 1.1 query language. W3C recommendation*. URL: <https://www.w3.org/TR/sparql11-query/>.
- He, X. and D. Golub (2016). “Character-level question answering with attention.” In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1598–1607. DOI: 10.18653/v1/D16-1166.
- Heer, J. and B. Shneiderman (2012). “Interactive dynamics for visual analysis: a taxonomy of tools that support the fluent and flexible use of visualizations.” In: *Queue* 10.2, pp. 30–55. DOI: 10.1145/2133416.2146416.
- Herrett, E., A. M. Gallagher, K. Bhaskaran, H. Forbes, R. Mathur, T. van Staa, and L. Smeeth (2015). “Data resource profile: Clinical Practice Research Datalink (CPRD).” In: *International Journal of Epidemiology* 44.3, pp. 827–836. DOI: 10.1093/ije/dyv098.
- Hertling, S., M. Schröder, C. Jilek, and A. Dengel (2016). “Top-k shortest paths in directed labeled multigraphs.” In: *Proceedings of the Third SemWebEval Challenge at ESWC 2016*, pp. 200–212. DOI: 10.1007/978-3-319-46565-4_16.
- Hilbert, D. M. and D. F. Redmiles (2000). “Extracting usability information from user interface events.” In: *ACM Computing Surveys* 32.4, pp. 384–421. DOI: 10.1145/371578.371593.
- Hirschman, L. and R. Gaizauskas (2001). “Natural language question answering: the view from here.” In: *Natural Language Engineering* 7.4, pp. 275–300. DOI: 10.1017/S1351324901002807.
- Hitzler, P., M. Krötzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph (2012). *OWL 2 Web Ontology Language primer (second edition)*. W3C recommendation. URL: <https://www.w3.org/TR/owl2-primer/>.
- Höffner, K., S. Walter, E. Marx, R. Usbeck, J. Lehmann, and A.-C. Ngonga Ngomo (2017). “Survey on challenges of question answering in the Semantic Web.” In:

Semantic Web 8.6, pp. 895–920. DOI: 10.3233/SW-160247.

Hripcsak, G., 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. Norén, Y.-C. Li, P. E. Stang, D. Madigan, and P. B. Ryan (2015). “Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers.” In: *Studies in Health Technology and Informatics* 216, pp. 574–578. DOI: 10.3233/978-1-61499-564-7-574.

Hripcsak, G., P. B. Ryan, J. D. Duke, N. H. Shah, R. W. Park, V. Huser, M. A. Suchard, M. J. Schuemie, F. J. Defalco, A. Perotte, J. M. Banda, C. G. Reich, L. M. Schilling, M. E. Matheny, D. Meeker, N. Pratt, and D. Madigan (2016). “Characterizing treatment pathways at scale using the OHDSI network.” In: *Proceedings of the National Academy of Sciences* 113.27, pp. 7329–7336. DOI: 10.1073/pnas.1510502113.

Hu, S., L. Zou, J. X. Yu, H. Wang, and D. Zhao (2018a). “Answering natural language questions by subgraph matching over knowledge graphs.” In: *IEEE Transactions on Knowledge and Data Engineering* 30.5, pp. 824–837. DOI: 10.1109/TKDE.2017.2766634.

Hu, S., L. Zou, and X. Zhang (2018b). “A state-transition framework to answer complex questions over knowledge base.” In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2098–2108. DOI: 10.18653/v1/D18-1234.

Huang, C., Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Z. Cheng, T. Yu, J. Xia, Y. Wei, W. Wu, X. Xie, W. Yin, H. Li, M. Liu, Y. Xiao, H. Gao, L. Guo, J. Xie, G. Wang, R. Jiang, Z. Gao, Q. Jin, J. Wang, and B. Cao (2020). “Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China.” In: *The Lancet* 395.10223, pp. 497–506. DOI: 10.1016/S0140-6736(20)30183-5.

Hyland, B., G. Ateazing, and B. Villazón-Terrazas (2014). *Best practices for publishing Linked Data. W3C working group note*. URL: <https://www.w3.org/TR/ld-bp/>.

- Jacobs, G., A. Wolf, M. Krawczak, and W. Lieb (2018). “Biobanks in the era of digital medicine.” In: *Clinical Pharmacology & Therapeutics* 103.5, pp. 761–762. DOI: 10.1002/cpt.968.
- Jacobsen, A., M. Thompson, M. Hanauer, B. Sergi, A. Gray, N. Juty, F. Ehrhart, C. Evelo, and M. Roos (2018). *D8.2: documentation of the tools for the data manipulation and standard conversions in the rare-disease field*. Report. ELIXIR-EXCELERATE. DOI: 10.5281/zenodo.1452467.
- Jin, H., Y. Luo, C. Gao, X. Tang, and P. Yuan (2019). “ComQA: question answering over knowledge base via semantic matching.” In: *IEEE Access* 7, pp. 75235–75246. DOI: 10.1109/ACCESS.2019.2918675.
- Kacprzak, E., L. M. Koesten, L.-D. Ibáñez, E. Simperl, and J. Tennison (2017). “A query log analysis of dataset search.” In: *Proceedings of the 17th International Conference on Web Engineering (ICWE)*, pp. 429–436. DOI: 10.1007/978-3-319-60131-1_29.
- Kaufmann, M., A. A. Manjili, P. Vagenas, P. M. Fischer, D. Kossmann, F. Färber, and N. May (2013). “Timeline index: a unified data structure for processing queries on temporal data in SAP HANA.” In: *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pp. 1173–1184. DOI: 10.1145/2463676.2465293.
- Kern, D. and B. Mathiak (2015). “Are there any differences in data set retrieval compared to well-known literature retrieval?” In: *Proceedings of the 19th International Conference on Theory and Practice of Digital Libraries (TPDL)*, pp. 197–208. DOI: 10.1007/978-3-319-24592-8_15.
- Khan, A., N. Li, X. Yan, Z. Guan, S. Chakraborty, and S. Tao (2011). “Neighborhood based fast graph search in large networks.” In: *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, pp. 901–912. DOI: 10.1145/1989323.1989418.
- Khurana, U. and A. Deshpande (2016). “Storing and analyzing historical graph data at scale.” In: *Proceedings of the 19th International Conference on Extending Database*

- Technology*, pp. 65–76. DOI: 10.5441/002/edbt.2016.09.
- Kitchenham, B. A., T. Dyba, and M. Jørgensen (2004). “Evidence-based software engineering.” In: *Proceedings of the 26th International Conference on Software Engineering (ICSE)*, pp. 273–281. DOI: 10.1016/B978-0-12-804206-9.00029-5.
- Knight, S.-a. and A. Spink (2008). “Toward a web search information behavior model.” In: *Web Search: Multidisciplinary Perspectives*. Springer. Chap. 12, pp. 209–234. DOI: 10.1007/978-3-540-75829-7_12.
- Knublauch, H. and D. Kontokostas (2017). *Shapes Constraint Language (SHACL)*. *W3C recommendation*. URL: <https://www.w3.org/TR/shacl/>.
- Köhler, S., 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. S. Dawkins, L. E. DeMare, A. D. Devereau, B. B. A. de Vries, H. V. Firth, K. Freson, D. Greene, A. Hamosh, I. Helbig, C. Hum, J. A. Jähn, R. James, R. Krause, S. J. F. Laulederkind, H. Lochmüller, 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. M. Veltman, T. Vulliamy, J. Yu, J. von Ziegenweidt, A. Zankl, S. Züchner, T. Zemojtjel, J. O. B. Jacobsen, T. Groza, D. Smedley, C. J. Mungall, M. Haendel, and P. N. Robinson (2016). “The Human Phenotype Ontology in 2017.” In: *Nucleic Acids Research* 45.D1, pp. D865–D876. DOI: 10.1093/nar/gkw1039.
- Kolker, E., E. Stewart, and V. Ozdemir (2012). “Opportunities and challenges for the life sciences community.” In: *OMICS: A Journal of Integrative Biology* 16.3, pp. 138–147. DOI: 10.1089/omi.2011.0152.
- Kulmanov, M., F. Z. Smaili, X. Gao, and R. Hoehndorf (2020). “Semantic similarity and machine learning with ontologies.” In: *Briefings in Bioinformatics* 22.4, pp. 1–18. DOI: 10.1093/bib/bbaa199.
- Lan, Y., S. Wang, and J. Jiang (2019). “Knowledge base question answering with a matching-aggregation model and question-specific contextual relations.” In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.10,

- pp. 1629–1638. DOI: 10.1109/TASLP.2019.2926125.
- Lancaster, O., T. Beck, D. Atlan, M. Swertz, C. Veal, R. Dalgleish, and A. Brookes (2015). “Cafe Variome: general-purpose software for making genotype-phenotype data discoverable in restricted or open access contexts.” In: *Human Mutation* 36.10, pp. 957–964. DOI: 10.1002/humu.22841.
- Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, P. Morsej Mohamed ans van Kleef, S. Auer, and C. Bizer (2015). “DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia.” In: *Semantic Web* 6.2, pp. 167–195. DOI: 10.3233/SW-140134.
- Lekschas, F. and N. Gehlenborg (2017). “SATORI: a system for ontology-guided visual exploration of biomedical data repositories.” In: *Bioinformatics* 34.7, pp. 1200–1207. DOI: 10.1093/bioinformatics/btx739.
- Li, G., P. Yuan, and H. Jin (2018). “Svega: answering natural language questions over knowledge base with semantic matching.” In: *Proceedings of the 30th International Conference on Software Engineering and Knowledge Engineering (SEKE)*, pp. 616–621. DOI: 10.18293/SEKE2018-119.
- Li, H., Y. Wang, S. Zhang, Y. Song, and H. Qu (2022). “*KG4Vis*: a knowledge graph-based approach for visualization recommendation.” In: *IEEE Transactions on Visualization and Computer Graphics* 28.1, pp. 195–205. DOI: 10.1109/TVCG.2021.3114863.
- Li, H., C. Xiong, and J. Callan (2017). “Natural language supported relation matching for question answering with knowledge graphs.” In: *Proceedings of the First Workshop on Knowledge Graphs and Semantics for Text Retrieval and Analysis (KG4IR)*, pp. 43–48. DOI: 10.1145/3132218.3132229.
- Li, W., A. Cowley, M. Uludag, T. Gur, H. McWilliam, S. Squizzato, Y. M. Park, N. Buso, and R. Lopez (2015). “The EMBL-EBI bioinformatics web and programmatic tools framework.” In: *Nucleic Acids Research* 43.W1, W580–W584. DOI: 10.1093/nar/gkv279.

- Lima, M. (2011). *Visual complexity: mapping patterns of information*. URL: <http://www.visualcomplexity.com/vc/book/>. New York: Princeton Architectural Press.
- Lloret-Gazo, J. (2016). “A survey on visual query systems in the web era.” In: *Proceedings of the 27th International Conference on Database and Expert Systems Applications (DEXA)*, pp. 343–351. DOI: 10.1007/978-3-319-44406-2_28.
- Lopes, P. and J. L. Oliveira (2013). “An innovative portal for rare genetic diseases research: the semantic Diseasecard.” In: *Journal of Biomedical Informatics* 46.6, pp. 1108–1115. DOI: 10.1016/j.jbi.2013.08.006.
- Lopez, V., P. Tommasi, S. Kotoulas, and J. Wu (2016). “QuerioDALI: question answering over dynamic and linked knowledge graphs.” In: *Proceedings of the 15th International Semantic Web Conference (ISWC)*, pp. 363–382. DOI: 10.1007/978-3-319-46547-0_32.
- Lukovnikov, D., A. Fischer, and J. Lehmann (2019). “Pretrained transformers for simple question answering over knowledge graphs.” In: *Proceedings of the 18th International Semantic Web Conference (ISWC)*, pp. 470–486. DOI: 10.1007/978-3-030-30793-6_27.
- Lukovnikov, D., A. Fischer, J. Lehmann, and S. Auer (2017). “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)*, pp. 1211–1220. DOI: 10.1145/3038912.3052675.
- Luo, D., J. Su, and S. Yu (2020a). “A BERT-based approach with relation-aware attention for knowledge base question answering.” In: *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. DOI: 10.1109/IJCNN48605.2020.9207186.
- Luo, K., F. Lin, X. Luo, and K. Zhu (2020b). “Knowledge base question answering via encoding of complex query graphs.” In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2185–2194. DOI: 10.18653/v1/D18-1242.

-
- Maali, F. and J. Erickson (2014). *Data Catalog Vocabulary (DCAT)*. *W3C recommendation*. URL: <https://www.w3.org/TR/vocab-dcat-1/>.
- Machanavajjhala, A., D. Kifer, J. Gehrke, and M. Venkatasubramanian (2007). “L-diversity: privacy beyond k-anonymity.” In: *ACM Transactions on Knowledge Discovery from Data* 1.1, pp. 1–52. DOI: 10.1145/1217299.1217302.
- Maheshwari, G., P. Trivedi, D. Lukovnikov, N. Chakraborty, A. Fischer, and J. Lehmann (2019). “Learning to rank query graphs for complex question answering over knowledge graphs.” In: *Proceedings of the 18th International Semantic Web Conference (ISWC)*, pp. 487–504. DOI: 10.1007/978-3-030-30793-6_28.
- Mahlmann, P. and C. Schindelhauer (2006). “Distributed random digraph transformations for peer-to-peer networks.” In: *Proceedings of the 18th Annual ACM Symposium on Parallelism in Algorithms and Architectures*, pp. 308–317. DOI: 10.1145/1148109.1148162.
- Marchionini, G. (2006). “Exploratory search: from finding to understanding.” In: *Communications of the ACM* 49.4, pp. 41–46. DOI: 10.1145/1121949.1121979.
- Martens, M., A. Ammar, A. Riutta, A. Waagmeester, D. N. Slenter, K. Hanspers, R. A. Miller, D. Digles, E. N. Lopes, F. Ehrhart, L. J. Dupuis, L. A. Winckers, S. L. Coort, E. L. Willighagen, C. T. Evelo, A. R. Pico, and M. Kutmon (2020). “WikiPathways: connecting communities.” In: *Nucleic Acids Research* 49.D1, pp. D613–D621. DOI: 10.1093/nar/gkaa1024.
- Marx, E., K. Höffner, S. Shekarpour, A.-C. N. Ngomo, J. Lehmann, and S. Auer (2016). “Exploring term networks for semantic search over RDF knowledge graphs.” In: *Proceedings of the 10th International Conference on Metadata and Semantics Research (MTSR)*, pp. 249–261. DOI: 10.1007/978-3-319-49157-8_22.
- Marx, E., A. Valdestilhas, H. Beck, and T. Soru (2021). “SANTé: A light-weight end-to-end semantic search framework for RDF data.” In: *The Semantic Web: ESWC 2021 Satellite Events*, pp. 93–97. DOI: 10.1007/978-3-030-80418-3_17.
- McNally, G., H. Rickards, M. Horton, and D. Craufurd (2015). “Exploring the validity of the short version of the Problem Behaviours Assessment (PBA-s) for

- Huntington's disease: a rasch analysis." In: *Journal of Huntington's Disease* 4.4, pp. 347–369. DOI: 10.3233/JHD-150164.
- McQuilton, P., A. Gonzalez-Beltran, P. Rocca-Serra, M. Thurston, A. Lister, E. Maguire, and S.-A. Sansone (2016). "BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences." In: *Database* 2016, pp. 1–8. DOI: 10.1093/database/baw075.
- Mishra, A. and S. K. Jain (2016). "A survey on question answering systems with classification." In: *Journal of King Saud University - Computer and Information Sciences* 28.3, pp. 345–361. DOI: 10.1016/j.jksuci.2014.10.007.
- Moher, D., A. Liberati, J. Tetzlaff, and D. G. Altman (2009). "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement." In: *BMJ* 339, pp. 1–8. DOI: 10.1136/bmj.b2535.
- Mons, B., C. Neylon, J. Velterop, M. Dumontier, L. O. B. d. Silva Santos, and M. D. Wilkinson (2017). "Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the European Open Science Cloud." In: *Information Services & Use* 37.1, pp. 49–56. DOI: 10.3233/ISU-170824.
- Morales, D. R., M. M. Conover, S. C. You, N. Pratt, K. Kostka, T. Duarte-Salles, S. Fernández-Bertolín, M. Aragón, S. L. DuVall, K. Lynch, T. Falconer, K. van Bochove, C. Sung, M. E. Matheny, C. G. Lambert, F. Nyberg, T. M. Alshammari, A. E. Williams, R. W. Park, J. Weaver, A. G. Sena, M. J. Schuemie, P. R. Rijnbeek, R. D. Williams, J. C. Lane, A. Prats-Urbe, L. Zhang, C. Areia, H. M. Krumholz, D. Prieto-Alhambra, P. B. Ryan, G. Hripcsak, and M. A. Suchard (2021). "Renin–angiotensin system blockers and susceptibility to COVID-19: an international, open science, cohort analysis." In: *The Lancet Digital Health* 3.2, e98–e114. DOI: 10.1016/S2589-7500(20)30289-2.
- Mulang, I. O., K. Singh, and F. Orlandi (2017). "Matching natural language relations to knowledge graph properties for question answering." In: *Proceedings of the 13th International Conference on Semantic Systems*, pp. 89–96. DOI: 10.1145/3132218.3132229.

- Nan, Y., J. D. Ser, S. Walsh, C. Schönlieb, M. Roberts, I. Selby, K. Howard, J. Owen, J. Neville, J. Guiot, B. Ernst, A. Pastor, A. Alberich-Bayarri, M. I. Menzel, S. Walsh, W. Vos, N. Flerin, J.-P. Charbonnier, E. van Rikxoort, A. Chatterjee, H. Woodruff, P. Lambin, L. Cerdá-Alberich, L. Martí-Bonmatí, F. Herrera, and G. Yang (2022). “Data harmonisation for information fusion in digital healthcare: a state-of-the-art systematic review, meta-analysis and future research directions.” In: *Information Fusion* 82, pp. 99–122. DOI: <https://doi.org/10.1016/j.inffus.2022.01.001>.
- Noy, N. and A. Rector (2006). *Defining n-ary relations on the Semantic Web. W3C working group note*. URL: <https://www.w3.org/TR/swbp-n-aryRelations/>.
- Ojokoh, B. and E. Adebisi (2018). “A review of question answering systems.” In: *Journal of Web Engineering* 17.8, pp. 717–758. DOI: 10.13052/jwe1540-9589.1785.
- Oliveira, J. L., A. Trifan, and L. A. B. Silva (2019). “EMIF Catalogue: a collaborative platform for sharing and reusing biomedical data.” In: *International Journal of Medical Informatics* 126, pp. 35–45. DOI: 10.1016/j.ijmedinf.2019.02.006.
- Panchbhai, A., T. Soru, and E. Marx (2020). “Exploring sequence-to-sequence models for SPARQL pattern composition.” In: *Proceedings of the 2nd Iberoamerican Knowledge Graphs and Semantic Web Conference (KGSWC)*, pp. 158–165. DOI: 10.1007/978-3-030-65384-2_12.
- Paraiso-Medina, S., D. Perez-Rey, R. Alonso-Calvo, B. Claerhout, K. de Schepper, P. Hennebert, J. Lhaut, J. Van Leeuwen, and A. Bucur (2013). “Semantic interoperability solution for multicentric breast cancer trials at the Integrate EU project.” In: *Proceedings of the 6th International Conference on Health Informatics (HEALTHINF)*, pp. 34–41. DOI: 10.5220/0004223400340041.
- Park, S., S. Kwon, B. Kim, and G. G. Lee (2015). “ISOFT at QALD-5: hybrid question answering system over Linked Data and text data.” In: *Proceedings of the 16th Conference and Labs of the Evaluation Forum (CLEF)*. URL: <http://ceur-ws.org/Vol-1391/127-CR.pdf>, pp. 1–11.
- Paulheim, H. (2017). “Knowledge graph refinement: a survey of approaches and evaluation methods.” In: *Semantic Web* 8.3, pp. 489–508. DOI: 10.3233/SW-160218.

- Peffer, K., T. Tuunanen, M. A. Rothenberger, and S. Chatterjee (2007). “A design science research methodology for information systems research.” In: *Journal of Management Information Systems* 24.3, pp. 45–77. DOI: 10.2753/MIS0742-1222240302.
- Penev, L., 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, and J. B. G. Lopez (2021). “Towards interlinked FAIR biodiversity knowledge: the BiCIKL perspective.” In: *Biodiversity Information Science and Standards* 5, pp. 1–3. DOI: 10.3897/biss.5.74233.
- Pereira, A., R. P. Lopes, and J. L. Oliveira (2020). “SCALEUS-FD: a FAIR data tool for biomedical applications.” In: *BioMed Research International* 2020, pp. 1–8. DOI: 10.1155/2020/3041498.
- Pereira, A., A. Trifan, R. P. Lopes, and J. L. Oliveira (2022). “Systematic review of question answering over knowledge bases.” In: *IET Software* 16.1, pp. 1–13. DOI: 10.1049/sfw2.12028.
- Perez-Riverol, Y., M. Bai, F. Leprevost, S. Squizzato, Y. Park, K. Haug, A. Carroll, D. Spalding, J. Paschall, M. Wang, N. Del Toro Ayllón, T. Ternent, P. Zhang, N. Buso, N. Bandeira, E. Deutsch, D. Campbell, R. Beavis, R. Salek, and H. Hermjakob (2017). “Discovering and linking public omics data sets using the Omics Discovery Index.” In: *Nature Biotechnology* 35.5, pp. 406–409. DOI: 10.1038/nbt.3790.
- Prud’hommeaux, E. and C. Buil-Aranda (2013). *SPARQL 1.1 federated query. W3C recommendation*. URL: <https://www.w3.org/TR/sparql11-federated-query/>.
- Reps, J., P. Ryan, P. Rijnbeek, and M. Schuemie (2021). “Design matters in patient-level prediction: evaluation of a cohort vs. case-control design when developing predictive models in observational healthcare datasets.” In: *Journal of Big Data* 8, pp. 1–18. DOI: 10.1186/s40537-021-00501-2.
- Reps, J. M., M. J. Schuemie, M. A. Suchard, P. B. Ryan, and P. R. Rijnbeek (2018). “Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data.” In:

-
- Journal of the American Medical Informatics Association* 25.8, pp. 969–975. DOI: 10.1093/jamia/ocy032.
- Rietveld, L. and R. Hoekstra (2013). “YASGUI: not just another SPARQL client.” In: *Proceedings of the ESWC2013 Workshop on Services and Applications over Linked APIs and Data*, pp. 78–86. DOI: 10.1007/978-3-642-41242-4_7.
- Robinson, I., J. Webber, and E. Eifrem (2015). *Graph databases, 2nd edition*. URL: <https://neo4j.com/lp/book-graph-databases/>. O’Reilly Media, Inc.
- Rodriguez, M. A. (2015). “The Gremlin graph traversal machine and language (invited talk).” In: *Proceedings of the 15th Symposium on Database Programming Languages (DBPL)*, pp. 1–10. DOI: 10.1145/2815072.2815073.
- Rodriguez-Iglesias, A., A. Rodríguez-González, A. G. Irvine, A. Sesma, M. Urban, K. E. Hammond-Kosack, and M. D. Wilkinson (2016). “Publishing FAIR data: an exemplar methodology utilizing PHI-base.” In: *Frontiers in Plant Science* 7, pp. 1–22. DOI: 10.3389/fpls.2016.00641.
- Rosenbaum, L. (2017). “Bridging the data-sharing divide - seeing the devil in the details, not the other camp.” In: *New England Journal of Medicine* 376.23, pp. 2201–2203. DOI: 10.1056/NEJMp1704482.
- Rücknagel, J., P. Vierkant, R. Ulrich, G. Kloska, E. Schnepf, D. Fichtmüller, E. Reuter, A. Semrau, M. Kindling, H. Pampel, M. Witt, F. Fritze, S. van de Sandt, J. Klump, H.-J. Goebelbecker, M. Skarupianski, R. Bertelmann, P. Schirnbacher, F. Scholze, C. Kramer, C. Fuchs, S. Spier, and A. Kirchhoff (2015). *Metadata schema for the description of research data repositories: version 3.0*. URL: https://gfzpublic.gfz-potsdam.de/pubman/item/item_1397899.
- Ruseti, S., A. Mirea, T. Rebedea, and S. Trausan-Matu (2015). “QAnswer - enhanced entity matching for question answering over Linked Data.” In: *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF)*. URL: <http://ceur-ws.org/Vol-1391/99-CR.pdf>, pp. 1–12.
- Salzberg, B. and V. J. Tsotras (1999). “Comparison of access methods for time-evolving data.” In: *ACM Computing Surveys* 31.2, pp. 158–221. DOI:

10.1145/319806.319816.

Sansone, S.-A., 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, and L. Ohno-Machado (2017). “DATS, the data tag suite to enable discoverability of datasets.” In: *Scientific Data* 4.1, pp. 1–8. DOI: 10.1038/sdata.2017.59.

Sansone, S.-A., P. McQuilton, P. Rocca-Serra, A. González-Beltrán, M. Izzo, A. Lister, and M. Thurston (2019). “FAIRsharing as a community approach to standards, repositories and policies.” In: *Nature Biotechnology* 37, pp. 358–367. DOI: 10.1038/s41587-019-0080-8.

Savenkov, D. and E. Agichtein (2016). “When a knowledge base is not enough: question answering over knowledge bases with external text data.” In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 235–244. DOI: 10.1145/2911451.2911536.

Sayers, E. W., E. E. Bolton, J. R. Brister, K. Canese, J. Chan, D. C. Comeau, R. Connor, K. Funk, C. Kelly, S. Kim, T. Madej, A. Marchler-Bauer, C. Lanczycki, S. Lathrop, Z. Lu, F. Thibaud-Nissen, T. Murphy, L. Phan, Y. Skripchenko, T. Tse, J. Wang, R. Williams, B. W. Trawick, K. D. Pruitt, and S. T. Sherry (2021). “Database resources of the national center for biotechnology information.” In: *Nucleic Acids Research* 50.D1, pp. D20–D26. DOI: 10.1093/nar/gkab1112.

Schaaf, J., D. Kadioglu, J. Goebel, C.-A. Behrendt, M. Roos, D. van Enckevort, F. Ückert, F. Sadiku, T. O. F. Wagner, and H. Storf (2018). “OSSE goes FAIR - implementation of the FAIR data principles for an open-source registry for rare diseases.” In: *Studies in health technology and informatics* 253, pp. 209–213. DOI: 10.3233/978-1-61499-896-9-209.

Scheider, S., A. Degbelo, R. Lemmens, C. van Elzakker, P. Zimmerhof, N. Kostic, J. Jones, and G. Banhatti (2017). “Exploratory querying of SPARQL endpoints in space and time.” In: *Semantic Web* 8.1, pp. 65–86. DOI: 10.3233/SW-150211.

Schmachtenberg, M., C. Bizer, and H. Paulheim (2014). “Adoption of the Linked Data best practices in different topical domains.” In: *Proceedings of the 13th*

-
- International Semantic Web Conference (ISWC)*, pp. 245–260. DOI: 10.1007/978-3-319-11964-9_16.
- Schneeweiss, S. and J. Avorn (2005). “A review of uses of health care utilization databases for epidemiologic research on therapeutics.” In: *Journal of Clinical Epidemiology* 58.4, pp. 323–337. DOI: 10.1016/j.jclinepi.2004.10.012.
- Schreiber, G. and Y. Raimond (2014). *RDF 1.1 primer. W3C Working Group note*. URL: <https://www.w3.org/TR/rdf11-primer/>.
- Schweiger, D., Z. Trajanoski, and S. Pabinger (2014). “SPARQLGraph: a web-based platform for graphically querying biological Semantic Web databases.” In: *BMC Bioinformatics* 15.1, pp. 1–5. DOI: 10.1186/1471-2105-15-279.
- Sequeira, M., J. R. Almeida, and J. L. Oliveira (2021). “A comparative analysis of data platforms for rare diseases.” In: *Proceedings of the 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 366–371. DOI: 10.1109/CBMS52027.2021.00041.
- Sernadela, P., L. González-Castro, C. Carta, E. v. d. 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 (2017a). “Linked registries: connecting rare diseases patient registries through a Semantic Web layer.” In: *BioMed Research International* 2017, pp. 1–13. DOI: 10.1155/2017/8327980.
- Sernadela, P., L. González-Castro, and J. L. Oliveira (2017b). “SCALEUS: Semantic Web services integration for biomedical applications.” In: *Journal of Medical Systems* 41.4, pp. 1–11. DOI: 10.1007/s10916-017-0705-8.
- Shekarpour, S., E. Marx, A.-C. Ngonga Ngomo, and S. Auer (2015). “SINA: semantic interpretation of user queries for question answering on interlinked data.” In: *Journal of Web Semantics* 30, pp. 39–51. DOI: <https://doi.org/10.1016/j.websem.2014.06.002>.
- Shen, W., J. Wang, and J. Han (2015). “Entity linking with a knowledge base: issues, techniques, and solutions.” In: *IEEE Transactions on Knowledge and Data*

- Engineering* 27.2, pp. 443–460. DOI: 10.1109/TKDE.2014.2327028.
- Siciliani, L., D. Diefenbach, P. Maret, P. Basile, and P. Lops (2019). “Handling modifiers in question answering over knowledge graphs.” In: *Proceedings of the 18th International Conference of the Italian Association for Artificial Intelligence (AIIA)*, pp. 210–222. DOI: 10.1007/978-3-030-35166-3_15.
- Siegler, E. L. (2010). “The evolving medical record.” In: *Annals of Internal Medicine* 153.10, pp. 671–677. DOI: 10.7326/0003-4819-153-10-201011160-00012.
- Silva, L. B., A. Trifan, and J. L. Oliveira (2018). “MONTRA: an agile architecture for data publishing and discovery.” In: *Computer Methods and Programs in Biomedicine* 160, pp. 33–42. DOI: 10.1016/j.cmpb.2018.03.024.
- Singh, K., A. Both, A. Sethupat, and S. Shekarpour (2018a). “Frankenstein: a platform enabling reuse of question answering components.” In: *Proceedings of the 15th European Semantic Web Conference (ESWC)*, pp. 624–638. DOI: 10.1007/978-3-319-93417-4_40.
- Singh, K., I. Lytra, M.-E. Vidal, D. Punjani, H. Thakkar, C. Lange, and S. Auer (2017). “QAestro - semantic-based composition of question answering pipelines.” In: *Proceedings of the 28th International Conference on Database and Expert Systems Applications (DEXA)*, pp. 19–34. DOI: 10.1007/978-3-319-64468-4_2.
- Singh, K., A. S. Radhakrishna, A. Both, S. Shekarpour, I. Lytra, R. Usbeck, A. Vyas, A. Khikmatullaev, D. Punjani, C. Lange, M. E. Vidal, J. Lehmann, and S. Auer (2018b). “Why reinvent the wheel: let’s build question answering systems together.” In: *Proceedings of the 27th International World Wide Web Conference (WWW)*, pp. 1247–1256. DOI: 10.1145/3178876.3186023.
- Song, D., F. Schilder, C. Smiley, C. Brew, T. Zielund, H. Bretz, R. Martin, C. Dale, J. Duprey, T. Miller, and J. Harrison (2015). “TR Discover: a natural language interface for querying and analyzing interlinked datasets.” In: *Proceedings of the 14th International Semantic Web Conference (ISWC)*, pp. 21–37. DOI: 10.1007/978-3-319-25010-6_2.

- Sorokin, D. and I. Gurevych (2017). “End-to-end representation learning for question answering with weak supervision.” In: *Proceedings of the 4th Semantic Web Evaluation Challenge (SemWebEval)*, pp. 70–83. DOI: 10.1007/978-3-319-69146-6_7.
- Speicher, S., J. Arwe, and A. Malhotra (2015). *Linked Data Platform 1.0. W3C recommendation*. URL: <https://www.w3.org/TR/ldp/>.
- Stang, P. E., P. B. Ryan, J. A. Racoosin, J. M. Overhage, A. G. Hartzema, C. Reich, E. Welebob, T. Scarnecchia, and J. Woodcock (2010). “Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership.” In: *Annals of Internal Medicine* 153.9, pp. 600–606. DOI: 10.7326/0003-4819-153-9-201011020-00010.
- Stearns, M., C. Price, K. Spackman, and A. Y. Wang (2001). “SNOMED clinical terms: overview of the development process and project status.” In: *Proceedings of the AMIA Annual Symposium*. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243297/>, pp. 662–666.
- Suchanek, F. M., G. Kasneci, and G. Weikum (2007). “YAGO: a core of semantic knowledge unifying WordNet and Wikipedia.” In: *Proceedings of the 16th International Conference on World Wide Web (WWW)*, pp. 697–706. DOI: 10.1145/1242572.1242667.
- Sweeney, L. (2002). “k-anonymity: a model for protecting privacy.” In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.5, pp. 557–570. DOI: 10.1142/S0218488502001648.
- Tagare, H. D., C. C. Jaffe, and J. Duncan (1997). “Medical image databases: a content-based retrieval approach.” In: *Journal of the American Medical Informatics Association* 4.3, pp. 184–198. DOI: 10.1136/jamia.1997.0040184.
- Tanon, T. P., D. Vrandečić, S. Schaffert, T. Steiner, and L. Pintscher (2016). “From Freebase to Wikidata: the great migration.” In: *Proceedings of the 25th International Conference on World Wide Web (WWW)*, pp. 1419–1428. DOI: 10.1145/2872427.2874809.

- Thabane, L., T. Thomas, C. Ye, and J. Paul (2009). “Posing the research question: not so simple.” In: *Canadian Journal of Anesthesia - Journal canadien d’anesthésie* 56.1, pp. 71–79. DOI: 10.1007/s12630-008-9007-4.
- Thompson, R., L. Johnston, D. Taruscio, L. Monaco, C. Bérout, 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, and H. Lochmüller (2014). “RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research.” In: *Journal of General Internal Medicine* 29.3, pp. 780–787. DOI: 10.1007/s11606-014-2908-8.
- Trifan, A. and J. L. Oliveira (2018). “A FAIR marketplace for biomedical data custodians and clinical researchers.” In: *Proceedings of the 31st International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 188–193. DOI: 10.1109/CBMS.2018.00040.
- Tsatsaronis, G., G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. Alvers, D. Weißenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artieres, A.-C. Ngonga Ngomo, N. Heino, E. Gaussier, L. Barrio-Alvers, and G. Paliouras (2015). “An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition.” In: *BMC Bioinformatics* 16, pp. 1–28. DOI: 10.1186/s12859-015-0564-6.
- UniProt Consortium (2020). “UniProt: the universal protein knowledgebase in 2021.” In: *Nucleic Acids Research* 49.D1, pp. D480–D489. DOI: 10.1093/nar/gkaa1100.
- Usbeck, R., R. Gusmita, M. Saleem, and A.-C. Ngonga Ngomo (2018). “9th challenge on question answering over linked data (QALD-9).” In: *Joint Proceedings of ISWC 2018 Workshops SemDeep-4 and NLIWOD-4*. URL: <http://ceur-ws.org/Vol-2241/paper-06.pdf>, pp. 58–64.
- Usbeck, R., A.-C. Ngonga Ngomo, L. Bühmann, and C. Unger (2015). “HAWK - hybrid question answering using Linked Data.” In: *Proceedings of the 12th European Semantic Web Conference (ESWC)*, pp. 353–368. DOI: 10.1007/978-3-319-18818-8_22.

- Vakulenko, S., J. D. Fernandez Garcia, A. Polleres, M. de Rijke, and M. Cochez (2019). “Message passing for complex question answering over knowledge graphs.” In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1431–1440. DOI: 10.1145/3357384.3358026.
- Vandenbussche, P.-Y., J. Umbrich, L. Matteis, A. Hogan, and C. Buil-Aranda (2017). “SPARQLES: monitoring public SPARQL endpoints.” In: *Semantic Web 8.6*, pp. 1049–1065. DOI: 10.3233/SW-170254.
- Villanueva, A. G., R. Cook-Deegan, B. A. Koenig, P. A. Deverka, E. Versalovic, A. L. McGuire, and M. A. Majumder (2019). “Characterizing the biomedical data-sharing landscape.” In: *Journal of Law, Medicine & Ethics* 47.1, pp. 21–30. DOI: 10.1177/1073110519840481.
- Vrandečić, D. and M. Krötzsch (2014). “Wikidata: a free collaborative knowledgebase.” In: *Communications of the ACM* 57.10, pp. 78–85. DOI: 10.1145/2629489.
- W3C SPARQL Working Group (2013). *SPARQL 1.1 overview. W3C recommendation*. URL: <https://www.w3.org/TR/sparql11-overview/>.
- Wade, T. D. (2014). “Traits and types of health data repositories.” In: *Health Information Science and Systems* 2.1, pp. 1–8. DOI: 10.1186/2047-2501-2-4.
- Wallis, J. C., E. Rolando, and C. L. Borgman (2013). “If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology.” In: *PLOS ONE* 8.7, pp. 1–17. DOI: 10.1371/journal.pone.0067332.
- Wang, R.-Z., Z.-H. Ling, and Y. Hu (2019). “Knowledge base question answering with attentive pooling for question representation.” In: *IEEE Access* 7, pp. 46773–46784. DOI: 10.1109/ACCESS.2019.2909826.
- Weinreich, S., R. Mangon, J. Sikkens, M. Teeuw, and M. Cornel (2008). “Orphanet: a european database for rare diseases.” In: *Nederlands Tijdschrift voor Geneeskunde* 152.9. URL: <https://pubmed.ncbi.nlm.nih.gov/18389888/>, pp. 518–519.
- Whetzel, P. L., N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen (2011). “BioPortal: enhanced functionality via new web services

- from the National Center for Biomedical Ontology to access and use ontologies in software applications.” In: *Nucleic Acids Research* 39.suppl_2, W541–W545. DOI: 10.1093/nar/gkr469.
- Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. O. Bonino 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, and B. Mons (2016). “The FAIR Guiding Principles for scientific data management and stewardship.” In: *Scientific Data* 3, pp. 1–9. DOI: 10.1038/sdata.2016.18.
- Wilkinson, M. D., S.-A. Sansone, E. Schultes, P. Doorn, L. O. Bonino da Silva Santos, and M. Dumontier (2018). “A design framework and exemplar metrics for FAIRness.” In: *Scientific Data* 5.1, pp. 1–4. DOI: 10.1038/sdata.2018.118.
- Wilkinson, M. D., R. Verborgh, L. O. Bonino da Silva Santos, T. Clark, M. A. Swertz, F. D. L. Kelpin, A. J. Gray, E. A. Schultes, E. M. van Mulligen, P. Ciccarese, A. Kuzniar, A. Gavai, M. Thompson, R. Kaliyaperumal, J. T. Bolleman, and M. Dumontier (2017). “Interoperability and FAIRness through a novel combination of web technologies.” In: *PeerJ Computer Science*, pp. 1–34. DOI: 10.7717/peerj-cs.110.
- Wylot, M., M. Hauswirth, P. Cudré-Mauroux, and S. Sakr (2018). “RDF data storage and query processing schemes: a survey.” In: *ACM Computing Surveys* 51.4. DOI: 10.1145/3177850.
- Xie, Z., Z. Zeng, G. Zhou, and T. He (2016). “Knowledge base question answering based on deep learning models.” In: *Proceedings of the 5th CCF Conference on Natural Language Processing and Chinese Computing (NLPCC) and 24th International Conference on Computer Processing of Oriental Languages (ICCPOL)*, pp. 300–311. DOI: 10.1007/978-3-319-50496-4_25.

- Xiong, W., M. Yu, S. Chang, X. Guo, and W. Y. Wang (2019). “Improving question answering over incomplete KBs with knowledge-aware reader.” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4258–4264. DOI: 10.18653/v1/P19-1417.
- Xu, K., Y. Feng, S. Huang, and D. Zhao (2014). “Question answering via phrasal semantic parsing.” In: *Proceedings of the 6th Conference and Labs of the Evaluation Forum (CLEF)*, pp. 414–426. DOI: 10.1007/978-3-319-24027-5_43.
- Xu, K., Y. Feng, S. Huang, and D. Zhao (2016a). “Hybrid question answering over knowledge base and free text.” In: *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*. URL: <https://aclanthology.org/C16-1226>, pp. 2397–2407.
- Xu, K., S. Reddy, Y. Feng, S. Huang, and D. Zhao (2016b). “Question answering on Freebase via relation extraction and textual evidence.” In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2326–2336. DOI: 10.18653/v1/P16-1220.
- Yamamoto, Y., A. Yamaguchi, and A. Splendiani (2018). “YummyData: providing high-quality open life science data.” In: *Database 2018*, pp. 1–12. DOI: 10.1093/database/bay022.
- Yi, J. S., Y. Kang, J. Stasko, and J. Jacko (2007). “Toward a deeper understanding of the role of interaction in information visualization.” In: *IEEE Transactions on Visualization and Computer Graphics* 13.6, pp. 1224–1231. DOI: 10.1109/TVCG.2007.70515.
- Yih, W.-t., M.-W. Chang, X. He, and J. Gao (2015). “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)*, pp. 1321–1331. DOI: 10.3115/v1/P15-1128.
- Yih, W.-t., M. Richardson, C. Meek, M.-W. Chang, and J. Suh (2016). “The value of semantic parse labeling for knowledge base question answering.” In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume*

- 2: *Short Papers*), pp. 201–206. DOI: 10.18653/v1/P16-2033.
- Yin, J., X. Jiang, Z. Lu, L. Shang, H. Li, and X. Li (2016). “Neural generative question answering.” In: *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 36–42. DOI: 10.18653/v1/W16-0106.
- Yin, P., N. Duan, B. Kao, J. Bao, and M. Zhou (2015). “Answering questions with complex semantic constraints on open knowledge bases.” In: *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pp. 1301–1310. DOI: 10.1145/2806416.2806542.
- Yu, M., W. Yin, K. S. Hasan, C. dos Santos, B. Xiang, and B. Zhou (2017). “Improved neural relation detection for knowledge base question answering.” In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 571–581. DOI: 10.18653/v1/P17-1053.
- Zafar, H., G. Napolitano, and J. Lehmann (2018). “Formal query generation for question answering over knowledge bases.” In: *Proceedings of the 15th European Semantic Web Conference (ESWC)*, pp. 714–728. DOI: 10.1007/978-3-319-93417-4_46.
- Zafar, H., G. Napolitano, and J. Lehmann (2019). “Deep query ranking for question answering over knowledge bases.” In: *Proceedings of the 18th Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pp. 635–638. DOI: 10.1007/978-3-030-10997-4_41.
- Zeshan, F., R. Mohamad, M. N. Ahmad, S. A. Hussain, A. Ahmad, I. Raza, A. Mehmood, I. Ulhaq, A. Abdulgader, and I. Babar (2017). “Ontology-based service discovery framework for dynamic environments.” In: *IET Software* 11.2, pp. 64–74. DOI: 10.1049/iet-sen.2016.0048.
- Zhang, H., G. Xu, X. Liang, G. Xu, F. Li, K. Fu, L. Wang, and T. Huang (2018). “An attention-based word-level interaction model for knowledge base relation detection.” In: *IEEE Access* 6, pp. 75429–75441. DOI: 10.1109/ACCESS.2018.2883304.
- Zheng, H.-T., Z.-Y. Fu, J.-Y. Chen, A. K. Sangaiah, Y. Jiang, and C.-Z. Zhao (2018a). “Novel knowledge-based system with relation detection and textual

evidence for question answering research.” In: *PLOS ONE* 13.10, pp. 1–21. DOI: 10.1371/journal.pone.0205097.

Zheng, W., J. X. Yu, L. Zou, and H. Cheng (2018b). “Question answering over knowledge graphs: question understanding via template decomposition.” In: *Proceedings of the 44th International Conference on Very Large Data Bases (VLDB)* 11.11, pp. 1373–1386. DOI: 10.14778/3236187.3236192.

Zheng, W., L. Zou, X. Lian, J. X. Yu, S. Song, and D. Zhao (2015). “How to build templates for RDF question/answering: an uncertain graph similarity join approach.” In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pp. 1809–1824. DOI: 10.1145/2723372.2747648.

