

University of Warsaw
Interdisciplinary Center for Mathematical
and Computational Modeling

Jakub Jałowiec

Student no. 358817

Evaluation of the *Linked Data*
approach to curation and processing
of data in epidemiology

Master's thesis
in COMPUTATIONAL ENGINEERING

The thesis was written under the supervision of:

Marek Michalewicz, Ph.D.

Interdisciplinary Center for Mathematical
and Computational Modeling

Piotr Bała, professor, Ph.D.

Interdisciplinary Center for Mathematical
and Computational Modeling

Warsaw, December, 2021

Abstract

This thesis aims to provide practical insights into the Linked Data approach to epidemiology. The insights are based on the preliminary results of a real-world study at the Pediatric Hospital of the Medical University of Warsaw which was concerned with investigating various medical and social risk factors on COVID-19 severity in children and SARS-CoV-2 spread dynamics. The thesis delivers results regarding the influence of various comorbidities on COVID-19 severity, based on a case of 90 patients of the Hospital, and establishes that metabolic disorders had a statistically significant influence on the hospitalization length in the observed sample ($\alpha = 0.05$). It discusses the role of Linked Data in obtaining that result and presents ways to reuse that approach to enable similar studies. An overview of the theoretical framework for the paradigm is presented and its practical implications are discussed. Additionally, the thesis provides a technical report of the implemented data processing pipeline.

Keywords

Linked Data, epidemiology, RDF, COVID-19, COVID-19 severity risk factors

Thesis domain (Socrates-Erasmus subject area codes)

11300 - Informatics, Computer Science

Subject classification

Life and medical sciences Health care information systems
Data management systems Information integration
Information systems Semantic web description languages

Tytuł pracy w języku polskim

Ocena podejścia *Linked Data* do przygotowania i przetwarzania danych epidemiologicznych

Acknowledgments

This thesis has been written as a part of the "Evaluation of the impact of the B.1.1.7 SARS-COV-2 virus variant on epidemiology and clinical picture of COVID-19 in children" study between the University of Warsaw (UW) and the Medical University of Warsaw (MUW).

I would like to express my sincerest gratitude to everyone involved in the study: Joanna Mańdziuk, Dr Magdalena Okarska-Napierała, Justyna Gadzińska, Weronika Woźniak, Krzysztof Piwoński, Dr Miron Kursa, Dr Aneta Afelt, Dr Katarzyna Suski-Grabowski and everyone in the administrative staff of the both universities that contributed to it. Without your work, writing this thesis would not have been possible.

Finally, I want to thank my supervisor, Dr Marek Michalewicz, for providing meritorical guidance throughout my studies.

Thank you all.

Contents

1. Introduction	7
1.1. The COVID-19 pandemic	7
1.2. The epidemiological study at the Pediatric Hospital of the Medical University of Warsaw	7
1.3. Goals & motivations of the thesis	8
1.4. Contributions	9
1.5. Related work	9
1.5.1. Linked Data in epidemiology	9
1.5.2. Severity of COVID-19 in children aged 0-18	10
1.6. Structure of the document	10
2. Methodology	11
2.1. Conceptual foundations of Linked Data	11
2.2. Designing the model	11
2.3. Implementation	12
2.4. Evaluation	12
3. Conceptual foundations of Linked Data	15
3.1. Linked Data	15
3.1.1. Background	15
3.1.2. Raw data vs. annotated data	16
3.1.3. The driving idea: linking heterogeneous data	17
3.2. Resource Description Framework	18
3.2.1. The core concept: triples	18
3.2.2. RDF as an abstract data model based on graphs	19
3.2.3. Syntax	21
3.2.4. Syntax vs. semantics	23
3.2.5. Semantics	25
3.3. The semantic stack of Linked Data	26
3.3.1. The four layers	27
3.3.2. Layer 0 – RDF as the basis	27
3.3.3. Layer 1 – Metamodels: RDFS & OWL	27
3.3.4. Layer 2 – Domain-specific ontologies	29
3.3.5. Layer 3 – Data integration	30
4. Results	31
4.1. Data model: the <i>covidepid</i> ontology	31
4.2. Data pipeline	32

4.3.	Risk factors of severe COVID-19 in children	34
4.3.1.	High-level overview of the data	34
4.3.2.	Hypothesis testing	34
5.	Discussion	37
5.1.	The role of Linked Data in the epidemiological study	37
5.2.	Statistical results	39
5.3.	Technical aspects	39
6.	Conclusions	41
6.1.	Linked Data in epidemiology	41
6.2.	Future work	42

Chapter 1

Introduction

According to World Health Organization, *coronavirus disease (COVID-19)* is an "*infectious disease caused by the SARS-CoV-2 virus*" [1]. Since the virus identification in late December 2019, the COVID-19 outbreak has had an immense impact on all aspects of everyday life. The disease most commonly manifests itself with a fever, cough, tiredness and loss of taste or smell [1]. In addition, it typically has a relatively high transmission risk and a few days-long latency phase.

1.1. The COVID-19 pandemic

In general, the primary route of spreading the virus is via direct person-to-person respiratory transmission where personal protective equipment is not used. Common circumstances of transmission include same-household contacts and healthcare settings like hospitals and long-term care facilities [2]. In addition, both children and adults can contract the virus, become symptomatic and infect others. Nevertheless, COVID-19 infections among children seem to be less severe and their mortality lower when compared to adults [3].

The World Health Organization estimates that as of the 10th December 2021, there have been over 267 million confirmed cases of COVID-19, including over 5 million deaths globally [4]. However, the COVID-19 pandemic is more than a health crisis. It profoundly affects societies and their economies, contributing to the shrinking of Gross Domestic Product (GDP) and a general increase in poverty and social inequalities [5]. Consequently, understanding how the disease spreads, introducing prevention and developing potential COVID-19 therapeutics have become global priorities.

1.2. The epidemiological study at the Pediatric Hospital of the Medical University of Warsaw

This thesis has been written as a part of an ongoing epidemiological study concerning the SARS-CoV-2 virus conducted at the Pediatric Hospital of the Medical University of Warsaw. The study started in December 2020 and will finish in summer 2022. The COVID-19 patients of the hospital and their families are interviewed in a retrospective, voluntary survey about their health and social conditions in order to provide insights about COVID-19 severity and spread dynamics in children aged 0-18.

Figure 1.1 presents phases of the epidemiological study. This thesis is concerned with data collection and analysis of the data obtained in the study's first phase, which spanned between December 2020 and July 2021.

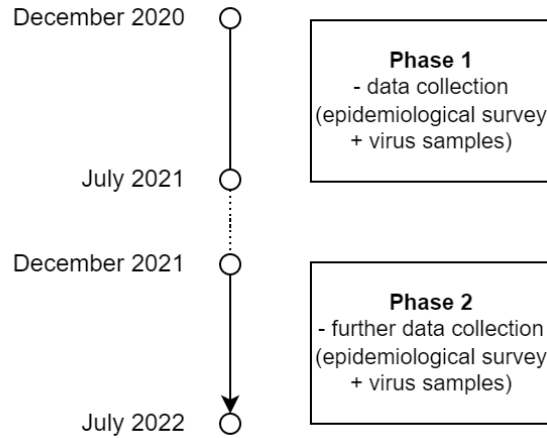


Figure 1.1: Phases of the epidemiological study.

1.3. Goals & motivations of the thesis

The ultimate goal of this thesis is to assess the usefulness of the *Linked Data* approach to data processing in epidemiology. *Linked Data* is a paradigm in data management proposed by World Wide Web Consortium [6]. It postulates to annotate heterogeneous datasets with a set of predefined annotations, called *Internationalized Resource Identifier (IRIs)*. *IRIs* are labels, usually taking form of HTTP addresses. The paradigm found its audience primarily in science, where data interoperability and comparability is especially crucial [7].

Prior to the SARS-CoV-2 pandemic, medicine had already been a subject of interest within the Linked Data community. As e.g. Kamdar et al. [8] argued in 2019, biomedicine suffered from the overwhelming plethora of data formats and lack of techniques to integrate them which hinders discoveries. They indicated pharmacology, cancer research and infectious diseases as branches of research which would gain through a widespread use of the Linked Data approach. They pointed out that further effort is needed to overcome difficulties which biomedical sciences face due to lack of common data management standards.

This thesis can be treated as a technical report of an application of the Linked Data-based approach to epidemiology. The two main goals of the thesis included:

1. design & implementation of a Linked Data processing system in the context of an epidemiological study, including:
 - (a) formulating a data model specific to the epidemiological study according to the Linked Data principles
 - (b) implementing a data pipeline conformant to the obtained data model
2. evaluation of the Linked Data approach to epidemiology by assessing risk factors of COVID-19 severity based on the data collected using the implemented system in the first phase of the epidemiological study

1.4. Contributions

As a result, a Linked Data model for the epidemiological study was formulated. The thesis presents some insights about the added value of that approach in the context of epidemiology. Additionally, it draws practical conclusions from the data collected in the first phase of the study. The main contributions of this thesis include:

1. a high-level overview of Linked Data-related concepts and notations, including RDF [9], Turtle [10], RDFS [11] and OWL [12] (Chapter 3.).
2. a technical report on the design (Section 4.1.) and implementation (Section 4.2.) of a data model in an epidemiological study formulated according to the Linked Data principles [13])
3. some insights about COVID-19 severity in children aged 0-18, obtained using the Linked Data approach (Section 4.3)

1.5. Related work

The following section presents the current state of the Linked Data paradigms and their practical applications to epidemiology. Additionally, some of the latest results in research on COVID-19 severity both in children and adults are provided.

1.5.1. Linked Data in epidemiology

The conceptual foundations of the paradigm were established in the works by Berners-Lee [14] [13] and by Berners-Lee et al. [15] [16] [17]. Allemang and Hendler [18] provide a good overview of notations used to representing Linked Data. Ultimately, Linked Data became the de-facto standard for scientific data management and stewardship [7].

With the beginning of the COVID-19 pandemic, applications of Linked Data in epidemiology have become a valuable area of research. Aakash et al. [19] as well as Bayoudhi et al. [20] provided surveys on Linked Data-related techniques for COVID-19 analytics, including a list of tools and various scenarios for their usage.

The results of Dutta et al. [21] are the most relevant for the epidemiological study at the Pediatric Hospital of the Medical University of Warsaw. They proposed the *CODO Ontology*: a conceptual model of *COVID-19 patients*, their *comorbidities*, *symptoms* etc. They reused the already available *SNOMED CT*, *FOAF* and *schema.org* vocabularies following the Linked Data principles. Unfortunately, their model lacks easy support for the ICD-10 vocabulary, which is used to report diagnosis and comorbidities in Polish hospitals. Additional mapping of their disease definition to the ICD-10 vocabulary would be needed. Although this can be achieved using e.g. the *Disease Ontology* [22], that effort would be out of the scope of this thesis. Additionally, the authors emphasize geographical annotations of the COVID cases, which is superfluous in the epidemiological study.

Further examples include e.g. the *CIDO* ontology by He et al. [23]. The authors proposed a general conceptual model for COVID-19 symptoms, disease transmission, and genetic conditions. The model heavily relies on the OBO ontology to make it as generic as possible. As a result, it proved to be too broad for the use case of the epidemiological study at the Pediatric Hospital of the Medical University of Warsaw.

1.5.2. Severity of COVID-19 in children aged 0-18

The following literature can be treated as a brief introduction to the topic of COVID-19 severity in children. Some references for COVID-19 hospitalization rates for various age groups and the outcomes of COVID-19 are presented. The retrospective studies provided here are examples of the most recent results in that area, both for adults and children. The typical medical and social risk factors analyzed in those studies included: comorbidities, obesity, pregnancy, age group, ethnicity, socioeconomic background.

The World Health Organization [3] indicated that the severity of COVID-19 tends to be lower in children than in adults. Thus, more investigation into severe COVID-19 cases in children needed to be made.

The study by Lindsay et al. [24] provided insights about the hospitalization rates for children at the beginning of the pandemic in the USA. The paper points out that most cases of COVID-19 among children seemed mild or moderate. At the same time, it stated that further effort should be put into confirming whether children are impacted differently by COVID-19 than adults. Delahoy et al. [25] in turn provided hospitalization rates per age group before and after the Delta variant became dominant in the USA based on the data obtained from the COVID-NET system in the USA [26].

Sandoval et al. [27] provided results for a cohort of 1853 COVID-positive young adults registered within the metropolitan healthcare system in Houston, Texas, including 226 pregnant women and 833 obese patients) whereas Woodruff et al. [28] for a sample of study sample of 454 children – patients of Children’s Hospital Colorado. The statistically significant results in both studies indicated respiratory system conditions, obesity, diabetes and preterm birth conditions as the risk factors of severe COVID-19. Rubenstein et al. [29] reported COVID outcomes of a sample of 82 children inpatients at three hospitals in the USA between spring and summer of 2020. The study points out the following observed risk factors of severe COVID-19: BMI above 25, higher age and comorbidities.

Other authors have also investigated the relationship between COVID-19 severity and specific health conditions, though primarily for adults. Examples include:

1. asthma: Gaietto et al. [30], Assaf et al. [31], Krishan et al. [32], Garcia-Pachon et al. [33]
2. pneumonia (Grandbastien et al. [34])

1.6. Structure of the document

Chapter 1. presents an introduction to the topic of Linked Data in the context of epidemiology – including its goals and motivations, as well as related work. Chapter 2. presents the methodology with regard to the design of a semantic data model as well as methods of statistical analysis, which were used to process the data collected in the epidemiological study. Chapter 3. provides a high-level overview of the relevant part of the Linked Data semantic stack, including descriptions of *RDF*, *RDFS*, *OWL* and some domain technologies. Chapter 4. presents results obtained in the thesis. Those included the semantic data model following the Linked Data principles, a report of technical aspects of the design and implementation of a semantic data model of COVID-positive patients in an epidemiological study and some statistical insights into the data collected in the first phase of the epidemiological study. Chapter 5. contains a discussion on the role of Linked Data in epidemiology. Chapter 6. provides conclusions and future work.

Chapter 2

Methodology

This thesis provides an evaluation of a practical application of the Linked Data approach in an epidemiological context. The leitmotif of the evaluation was to assess risk factors on COVID-19 severity in children aged 0-18 admitted to the Pediatric Hospital between December 2020 and July 2021. In the following sections, a description of the methodology chosen in this work follows.

2.1. Conceptual foundations of Linked Data

A deeper overview of the conceptual foundations of Linked Data can be found in Chapter 3. It involved:

1. reference & citation analysis in the following area:
 - (a) classical works on Linked Data, including those done by Berners-Lee [13] and Berners-Lee et al. [17]
 - (b) specifications of the technologies related to data storage and modeling using Resource Description Framework [9] and its associated metadata modeling languages: Resource Description Framework Schema [11], Web Ontology Language [12]
 - (c) ontologies and their applications in the context of domain modeling, including classical works on ontologies by Gruber [35] and on domain-specific ontologies such as ICD10CM [36], ATC [37] and FOAF [38]
2. formalization of the Resource Description Framework's abstract syntax in terms of:
 - Extended Backus Naur Form [39, Chapter 6. Notation] (Definition 3.2.2.)
 - graph theory [40] (Definition 3.2.1.)

2.2. Designing the model

A major challenge of the epidemiological study at the Pediatric Hospital of the Medical University of Warsaw was to design a data-processing methodology that would enable seamless integration of data of various provenance, including medical, pharmaceutical and socio-economic vocabularies.

One of the data-processing paradigms which help to fulfill these requirements is Linked Data [13]. As opposed to "traditional" approaches, Linked Data rigorously defines the semantics of the data [41], using *Resource Description Framework* and *ontologies* [18] as its formal foundations [9] [41].

The semantic model of COVID-positive patients in the epidemiological study at the Pediatric Hospital of the Medical University of Warsaw was designed according to the principles of Linked Data [13]. It involved linking two different datasets of biomedical terms: ICD-10CM [36] and ATC [37] with a vocabulary of concepts related to *humans*: FOAF [38].

The design of the model was conducted using:

1. Protege v5.5.0 [42] – an RDF-based data model editor
2. BioPortal (<https://bioportal.bioontology.org/>) [43] – a search engine for various RDF-based vocabularies in the context of biology and medicine

2.3. Implementation

Figure 2.1. presents an overview of the data pipeline implemented for the obtained data model. The data pipeline was deployed on a Linux machine, using Docker [44].

The pipeline consisted of four steps:

1. manual entry of surveys through a user interface
2. mapping of the JSON data to an RDF representation
3. bulk data load into Apache Jena [45], a tool capable of applying inferences to Linked Data
4. statistical analysis and Apache Jena querying within RStudio [46]

A more detailed report on the technical implementation of the data pipeline can be found in Section 4.2.

2.4. Evaluation

The ultimate goal of this thesis was to assess the usefulness of the Linked Data approach in the context of an epidemiological study. The assessment was based on the ability to answer the following analytical question: what comorbidities impacted COVID-19 severity in the obtained sample? Patient's hospitalization was selected as the indicator of COVID-19 severity.

The data were verified to have a non-normal distribution using the Shapiro-Wilk test. It was assumed that the presence of comorbidities positively influenced COVID-19 severity. Thus, the one-sided version of Mann-Whitney-Wilcoxon's U test was used to test whether cases *with* comorbidity had a longer hospitalization length. In order to achieve a proper level of data granularity, the comorbidities were grouped by their class (e.g. *respiratory system diseases*, *neurological diseases*) according to the ICD-10CM taxonomy [36]. Comorbidities falling into the same group were counted separately for a single patient.

In order to prevent data noise and underrepresented comorbidity classes from being recognized as risk factors, the cut-off value for their observed count was set to min. 10% of the

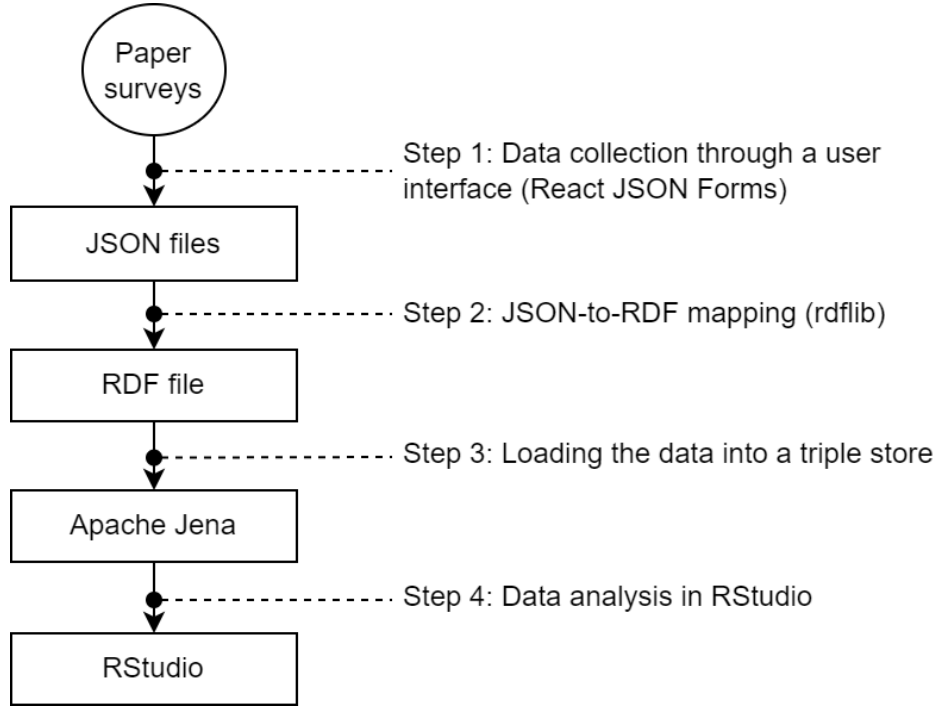


Figure 2.1: Data processing pipeline. The data were transformed from surveys collected on paper to a Linked-Data-based representation in Turtle [10], available for processing in RStudio [46].

total population. Otherwise the comorbidity class was rejected from further investigation as a possible risk factor.

Table 2.1 presents the general form of the hypothesis under investigation. It was tested whether the distributions of both populations are equal.

H_0 hypothesis	$P(X > Y) = 0.5 - (X, Y) \in (S_{true}, S_{false})$
H_1 hypothesis	$P(X > Y) > 0.5 - (X, Y) \in (S_{true}, S_{false})$

Table 2.1: The general scheme of the investigated hypotheses. Given two subsamples S_{true} (comorbidity class present) and S_{false} (comorbidity class absent) it was checked whether for any two values $(X, Y) \in (S_{true}, S_{false})$, the probability that the hospitalization length of a person with the comorbidity (X) was greater than that of a person without the comorbidity (Y) is higher than 0.5.

Chapter 3

Conceptual foundations of Linked Data

As pointed out in Section 1.3., Linked Data has been chosen as the data processing paradigm for the epidemiological study at the Pediatric Hospital of the Medical University of Warsaw. The following chapter provides an overview of the conceptual foundations of the Linked Data paradigm. It presents both theoretical as well as practical aspects of the Linked Data technology stack.

3.1. Linked Data

The Linked Data paradigm has been postulated by Tim Berners-Lee [13], one of the inventors of the *World Wide Web* [15]. It aims to make data of various provenance machine interpretable and reusable. It gained popularity especially in the area of scientific data management and stewardship [7].

Linked Data uses *Resource Description Framework (RDF)* [9] [13] as its abstract data model. It reuses its concept of annotating the data using *Internationalized Resource Identifiers (IRIs)*. IRIs provide an infrastructure for linking various data sources and facilitate *shared semantics* for them. The ultimate goal of Linked Data is to provide infrastructure for data integration through reuse of IRIs in various contexts. Throughout this thesis, the term Linked Data will be used as a synonym for the term *Semantic Web* to denote a distributed *knowledge graph* labeled with IRIs [47].

3.1.1. Background

The need to provide a formal yet simplified way to annotate data has its roots in the so-called *AAA problem* [9] [14] [18], which stands for *Anybody can say Anything about Anything*. The problem boils down to the following observations about modern data that most analysts face:

1. huge amounts of data are made publicly available by various organizations on the Internet, such as public services, scientists, researchers, governments and private companies, which want to share their data
2. the data available publicly often uses "ad-hoc" semantics, which prevents their easy processing outside of the context the data were published in

Tim Berners-Lee [13] proposed four principles that the data published on the Internet should follow in order to be automatically interpretable by computers [13], thus facilitating its analysis. He called them the "Linked Data principles":

1. all *entities* (such as *study subjects*, *diseases*, *medical substances*, *measurements* etc.) and *types of relationships* (e.g. "*measurement has value*", "*person has diseases*") within the frame of discourse should be assigned so called *Internationalized Resource Identifiers (IRI)* – labels that globally identify those *things* (also called *resources*)
2. *IRIs* which are also *HTTP addresses* should be viewable through Internet browsers in order to look up their meaning by human users
3. *IRIs* should provide machine-interpretable interpretation of the concepts they represent
4. when possible, all data should reuse existing *IRIs* in order to facilitate shared meaning and let the referer discover other *things* (*resources*) by looking up the other *IRIs* ("following the links")

Although those principles were initially formulated for the Internet (Berners-Lee proposed them as a way to structurize the World Wide Web in terms of data semantics) and found little audience in the web development community, they provided important conceptual foundations for annotating *any* kind of data with *IRIs*. In fact, we will treat any data annotated with *IRIs* as Linked Data, regardless whether the *IRI* point to a place on the Internet. As such, *IRIs* can be treated as plain labels with which data can be annotated to provide meaning for them.

Development of new databases and programs that support Linked Data processing, such as *databases* (also called *triple stores* in the community), and *inference engines* (*reasoners*) is an active area of research for *semantic* data science. The *SPARQL Protocol and RDF Query Language* [48] provides both a query language and APIs to access that data programmatically. The technological stack of Linked Data constitutes a well-established alternative to the already existing *semantics-agnostic* data stores, such as SQL databases [49] or plain files, e.g. in JSON or XML [50] format.

3.1.2. Raw data vs. annotated data

One of the main goals of the Linked Data paradigm is to add a semantic layer to *any* data [13]. It facilitates its *shared meaning* through the use of special *annotations* called *Internationalized Resource Identifiers (IRIs)*.

Table 3.1 presents a small example of how annotations using the Internationalized Resource Identifiers work in practice. We will treat all labels present in the b) subfigure but not present in the a) subfigure as *IRIs*.

The main problem with the data in the a) subfigure in Table 3.1. is that there is no obvious way of saying what that data actually means. The naming convention of the columns ("*id*" on the left and "*cond*" on the right) alludes to *identifiers* and some sorts of *conditions*, only the person knowing the use case and storage format understands their meaning.

On the other hand, the b) subfigure presents the same data *with annotations*. Through their use, the data becomes (more or less) human-readable – the presented example models human's *comorbidities*. It can trivially be concluded that *N9X2*, *K1A4* and *P0Z2* are identifiers of some *persons* and that *U07.1*, *J45.1* and *G40.909* are codes of *health conditions*. In fact, it is easy to check that those codes correspond in the ICD-10 vocabulary [36] to the

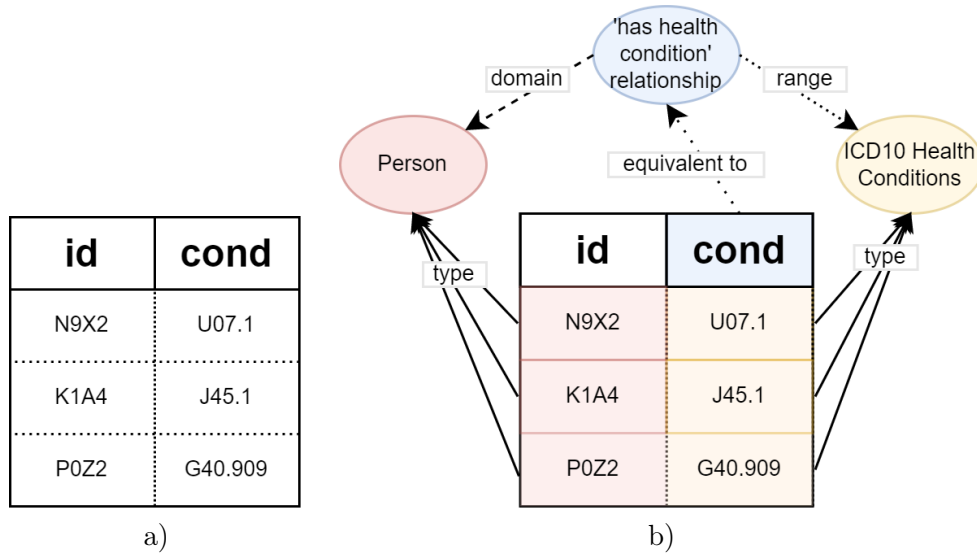


Table 3.1: *Raw data* (left) vs. *data with annotations* (right). The example in the b) subfigure conceptually follows the Linked Data approach (it provides explicit semantics for the data through annotations) whereas the example in the a) subfigure does not – it only has some implicit semantics for the owner of the data.

COVID-19 disease, *asthma* and *epilepsy*, respectively. Moreover, those annotations can be expressed using a formal notation – such as RDF (see Chapter 3.) – in order to make them also machine interpretable [9].

The philosophical considerations of whether it is possible to create a schema-neutral system of annotations using IRIs and assessing its usefulness has been a topic of a long-term debate within the Linked Data community. Doctorow [51] and Schwartz [52] provide two different views on that matter. Regardless of their views, this thesis treats the annotations using IRIs as a practical way to add any sort of semantics to the data.

3.1.3. The driving idea: linking heterogeneous data

As argued above, annotations provide meaning to the data. Moreover, they can be written in a formal notation (such as RDF [9].), which allows them to be treated as any other data. Using annotations makes the data interpretable regardless of the platform and methodology of data processing – as long as it can interpret those annotations.

At its core, the Linked Data paradigm tries to achieve data interoperability in two steps:

1. by publishing custom annotation types formally describing their meaning by the users (e.g. using RDF [9])
2. more importantly, reusing annotations created by others in your own datasets

Linked Data especially emphasizes the second point: reusing annotations of others and thus interlinking datasets – hence the name. In fact, provision and curation of those links constitutes the major part of activities around Linked Data.

It is necessary to discuss the practical implications of the second point for the b) subfigure of Table 3.1. The presented hypothetical dataset provides data about *persons* and their *comorbidities*. At the same time, consider another hypothetical dataset which links *diseases* to

physiological systems of human body they affect (such as *neurological system*, *immunological system* etc.). If those datasets used the same taxonomies for diseases, then their integration would be trivial. In other words, investigation whether comorbidities of certain systems in human body influence COVID-19 severity would be easy to conduct.

Annotations of various datasets using IRIs is the core of Linked Data. Providing them wherever possible guarantees seamless data integration in the future.

3.2. Resource Description Framework

As pointed out in Section 3.1.2., *Resource Description Framework* brings formal framework for syntax and semantics of annotations using IRIs. Thus, it provides foundations for Linked Data. According to its authors, "*RDF* is an assertional language intended to be used to express propositions using precise formal vocabularies" [41]. In fact, it can be regarded as general-purpose language to express statements about any data. The following section presents a formal definition of RDF syntax and an informal description of RDF semantics.

The current standard is defined in the following list of documents [53]:

- *RDF 1.1 Concepts and Abstract Syntax* [9]
- *RDF 1.1 Semantics* [41]
- a number of serialization specifications, e.g.: *RDF 1.1 XML Syntax* [54] and *RDF 1.1 Turtle* [10]

3.2.1. The core concept: triples

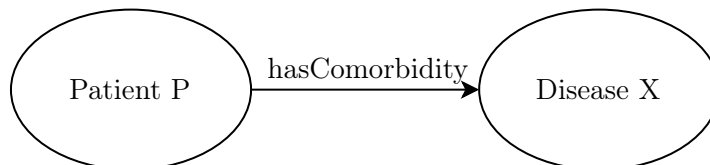


Figure 3.1: An example of an *RDF triple* represented as a graph. *Patient P* is the *subject*, *hasComorbidity* is the *predicate* and *Disease X* is the *object*. The direction of the arc in the graphical notation is important, i.e. the relationship is not necessarily symmetrical.

The so called *triples* are the central concept of the Resource Description Framework. A triple consists of a *subject*, a *predicate* and an *object* [9]. They provide a generic way of making claims about the universe of discourse. An example of a triple is the sentence "*Patient P has comorbidity Disease X*", where *Patient P* is the *subject*, *has comorbidity* is the *predicate* and *Disease X* is the *object*. Figure 3.1 shows a graphical representation of the example triple.

A triple denotes a single statement about the *world*. All such statements are collectively called *assertions*. Assertions that are objectively proven to be true are called *facts*. Boolean valuation of facts is either provided by a real-world observation (such as an experiment) or by derivation from other true assertions through the means of *logical inference* (refer to Section 3.2.5.).

Examples of various assertions are provided below:

- *Patient P has comorbidity Y*

- *All patients were COVID-positive*
- *Comorbidity Z could influence COVID-19 severity*
- *Person Q claims to have got infected in work*

The first sentence is a simple, database-like statement about a single observation. The second sentence is an example of quantification. The third sentence is a claim which is a-priori unknown to be false or true. The fourth sentence is a claim about another sentence. All those sentences have various degrees of abstraction (simple fact vs. quantification vs. uncertainty vs. statements about statements) and can be easily expressed in RDF.

3.2.2. RDF as an abstract data model based on graphs

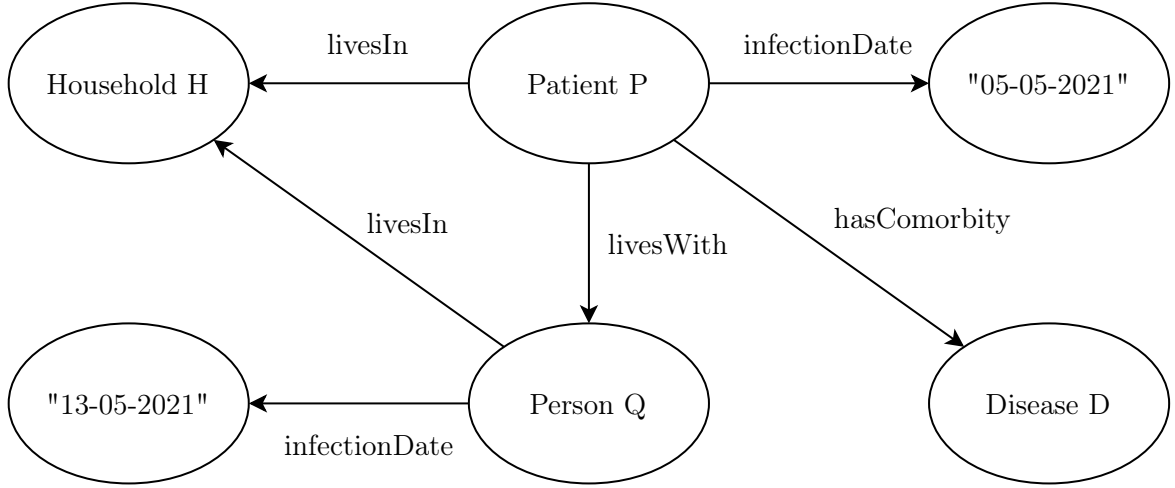


Figure 3.2: An example of an *RDF graph*. *Things* in the universe of discourse ("*resources*") are drawn as ovals (also called *nodes*). *Arcs* (also called *edges*) are labeled with *relationships*. There are 6 edges which means that this RDF graph contains 6 triples (assertions).

The *RDF* specification often refers to an *RDF graph* as its driving concept. Generally speaking, a *graph* (in some contexts called also a *network*) is a mathematical structure, which models relationships between *entities* of the universe of discourse. *Graphs* have a well-established mathematical theory behind them which has its roots in the works of Euler [55] and their properties are relatively well studied and understood. Kenneth and Wright [40] provided a formal definition of *graphs* with proofs of their mathematical properties and some associated algorithms.

Roughly speaking, graph's primary usage is to express the *abstract structure* of a given problem. A widely used, intuitive graphical notion for them involves drawing *circles* connected by *arcs*. They are used in those branches of research where *the structure* carries inherent value, e.g. in *transportation networks*, *social networks*, *electronic circuits* and *topology*. One of the major applications of graphs is *metamodeling*. *Entity-relationship models* and *taxonomies* are all expressed using graphs.

Ehrlinger et al. [56] provide a definition of a *knowledge graph*, that is a graph used specifically in the context of knowledge representation. Resource Description Framework can be regarded as a notation to represent knowledge graphs.

The RDF specification often mixes different naming conventions from graph theory, description logics and computer science. We will uniformly refer to the elements modeled in graphs as *nodes* or *resources* and to the links between them as *relationships*. Both *nodes* and *relationships* do not have any internal structure – they are plain labels.

Nodes in RDF come in two flavors:

1. *concepts* – they model entities in the domain of the discourse
2. *literals* – they represent atomic data values, such as *numbers*, *dates*, *strings* etc.

The purpose of edges in RDF is to model relations between nodes.

Definition 3.2.1 proposes a formal notation of an *RDF graph*. The sets of labels R_i , R_b and R_l are the building blocks of an *RDF graph*. They contain the so called *IRIs*, *blank nodes* and *literals*, respectively, collectively known as *resources*. As pointed out in Section 3.1.1., *IRIs* are used to globally identify resources. *Blank nodes* in turn are used to identify things within the scope of a single *RDF graph*. *Literals* are used to denote values of atomic types, such as *strings*, *numbers*, *booleans*, *dates*. The S , P , O letters denote the sets of *subjects*, *predicates* and *objects*.

Definition 3.2.1 (RDF graph) *An RDF graph is a tuple (S, P, O, R_i, R_b, R_l) where:*

1. R_i, R_b, R_l – three disjunctive finite sets of labels called *Internationalized Resource Identifiers (IRIs)*, *blank node identifiers* and *literals* respectively; the set $R_i \cup R_b \cup R_l$ is called "*resources*"
2. $S \subseteq R_i \cup R_b$ is a finite set of labels called *subjects*
3. $O \subseteq R_i \cup R_b \cup R_l$ is a finite set of labels called *objects*
4. $P \subseteq S \times O \times R_i$ is a finite set of tuple called *properties*

As can be seen:

- a *subject* can be either an *Internationalized Resource Identifiers* or a *literal value*
- a *predicate* can only be an *Internationalized Resource Identifiers*
- an *object* can be either an *Internationalized Resource Identifiers*, a *literal value* or a *blank node identifier*

Thus, it is possible for an IRI to appear in some triples as a *node* (either a *subject* or an *object*) and in other triples as a *predicate*. That does not however lead ultimately to an inconsistency but it is just a construct available in RDF to express *statements* about *statements*.

Figure 3.2 shows a graphical interpretation of an *RDF graph*. *Nodes* are drawn as the rounded shapes with their designating labels inside them. *Relationships* are drawn as arrows starting from one node to another. A *relationship* (s, t, l) is drawn as an arrow from the node s (also called the *source*) towards the node t (called the *target*) annotated with the label l . The (s, t) pair can be regarded as the *direction* of the *edge* and in general it is a binary relationship determining connectivity between nodes. Thus, *edges* are directed and, unless explicitly stated, $(s, t, l) \in P$ does not imply $(t, s, l) \in P$.

3.2.3. Syntax

The authors of RDF provided a formalization of how RDF data should be written, called *RDF abstract syntax*. It can be regarded as a minimal storage format for RDF data. Actual implementations of the abstract syntax are called *concrete syntaxes*. There is a multitude of concrete syntaxes available, which can lead to confusion when using RDF in general. The following sections discuss the most important aspects of RDF syntax.

Abstract syntax

The RDF's *abstract syntax* can be defined using the Extended Backus-Naur Form notation¹.

Definition 3.2.2 (RDF abstract syntax) *RDF abstract syntax is a syntax whose rules are described by Listing 3.1.*

```
data          ::= triple*
triple        ::= ws* subject ws+ predicate ws+ object ws+ separator
subject       ::= iri | blankNodeID
predicate     ::= iri
object        ::= iri | blankNodeID | literal
```

Listing 3.1: Abstract syntax of Resource Description Framework defined using the *EBNF* notation (refer to Table 3.2.).

The following symbols were left undefined in the Listing 3.1 as the RDF specification does not necessarily enforce any particular representation for them:

- `ws` – any symbol designating a single whitespace character, such as *space*, *new line* etc.
- `separator` – a single symbol terminating a triple definition
- `iri` – a symbol designating an *Internationalized Resource Identifier* (e.g. a HTTP address)
- `blankNodeID` – a symbol designating a *blank node*
- `literal` – a symbol designating a *literal values*

As can be seen, the abstract syntax of the *Resource Description Framework* mimics the definition of an *RDF graph*. The following concepts from the definition of an RDF graph (Definition 3.2.1.) are present in it: *triple*, *subject*, *predicate*, *object*, *IRIs*, *blank nodes*, *literals*. Each edge in an RDF graph is mapped to a separate triple.

¹<https://www.w3.org/TR/REC-xml/#sec-notation>

aaa, bbb etc.	syntactical symbols
::=	substitution equivalence (left side can be substituted by right side)
aaa bbb	concatenation of two symbols
aaa bbb	alternative between two symbols
aaa*	zero or more occurrences of the symbol (Kleene star)
aaa+	one or more occurrences of the symbol

Table 3.2: A brief summary of the subset of *Extended Backus-Naur Form* used to define the abstract syntax of RDF.

Abstract vs. concrete syntax

Aside from defining syntax for *whitespaces*, *triple separators*, *IRIs*, *blank node identifiers* and *literals*, the abstract syntax leaves two things to consider when designing a *concrete syntax*:

1. optimization of the resulting file size
2. human readability of the concrete syntax

Trivial ways to optimize the resulting file size and at the same time to improve its human readability include using shorthands for IRI prefixes and listing triples grouped by subject.

The following comparison tries to justify the need to thoroughly implement a syntax for RDF. Consider two hypothetical representations of the same data: one represented in the *abstract syntax way*, where each triple is in a separate line (Listing 3.2) and one optimized using the postulated improvements (Listing 3.3). Clearly, in the latter case the readability has been improved as well as the size of the resulting data file decreased.

```
http//<url>#patient_P http//<url>#livesWith http//<url>#person_Q .
http//<url>#patient_P http//<url>#livesIn http//<url>#household_H .
http//<url>#patient_P http//<url>#hasComorbidity http//<url>#disease_D .
```

Listing 3.2: Strict use of the RDF abstract syntax to store triples.

```
@prefix ns: <http//<url>#> .
ns:patient_P ns:livesWith ns:person_Q ;
              ns:livesIn ns:household_H ;
              ns:hasComorbidity ns:disease_D ;
```

Listing 3.3: The same data as in Listing 3.2. represented using an optimized *concrete syntax*. Compared to the other version grouping triples by subjects and extracting IRI prefixes decreased the resulting file size (the text is shorter) and improved its readability.

Unfortunately, there is no single, commonly used standard *concrete syntax*. The authors of RDF provided three different concrete syntaxes (*RDF/XML* [54], *JSON-LD* [57], *Terse RDF Triple Language* [10]) and many more unofficial standards emerged from the RDF community. Thus, RDF programs are forced to support multiple different input formats at the same time. It leads to unnecessary confusion between users across different formats. A partial solution is to use RDF format converters, such as Apache Jena’s *riot* [45].

An example of a concrete syntax: Turtle

To wrap up, let’s consider one example of RDF concrete syntaxes: the *Terse RDF Triple Language* [10] (*Turtle*). Listing 3.4 presents the example RDF graph from Figure 3.2 represented in that format. In fact, a similar example has already been seen in the Listing 3.2.


```

@prefix covidpid: <https://github.com/kubajal/covidpid#> .
covidpid:patient_P covidpid:livesWith covidpid:person_Q ;
                  covidpid:livesIn covidpid:household_H ;
                  covidpid:hasComorbidity covidpid:disease_D ;
                  covidpid:infectionDate "05-05-2021" .
covidpid:person_Q covidpid:livesIn covidpid:household_H ;
                  covidpid:infectionDate "13-05-2021" .

```

Listing 3.4: The graph from Listing 3.4. stored using the *Turtle* format. There are two subjects: patient_P and patient_Q four different predicates: livesWith livesIn hasComorbidity infectionDate and five different subjects: person_Q household_H disease_D "05-05-2021" "13-05-2021".

The example from Listing 3.4 should be read in the following way:

- patient_P lives with person_Q
- patient_P lives in household_H
- patient_P has comorbidity disease_D
- patient_P was infected on the 5th of Mai 2021
- person_Q lives in household_H
- person_Q was infected on the 13th of Mai 2021

Notable features of the *Turtle* concrete syntax used in the Listing 3.4. include:

- declaration of prefixes as shorthands for RDF namespaces (*@prefix <prefix> <namespace>*)
- access to identifiable resources (such as patient_p) through the *<prefix>: patient_X*
- separation of (*predicate object*) pairs for the same *subject* by the ";" character

3.2.4. Syntax vs. semantics

Until now, it has been stated how RDF can be used to write down data using various notations based on its abstract syntax (such as Turtle) and that its underlying data model is in fact a graph. The expressivity of RDF was not an issue whatsoever as we have treated all resources and relationships as pure labels that had only *implicit* meaning to us.

Assigning meaning to labels

```
covidpid:patient_X covidpid:infectionDate "05-05-2021" .
```

Listing 3.5: An example of an assertion.

Listing 3.5 provides a driving example in this subsection. Without the context of IRIs, the asserted triple implicitly means that "*patient_X* was infected on the 5th of May", including that e.g.:

1. *patient_X* is probably an identifier of a *human person*

2. the domain of the *infectionDate* relationship are *infected persons* and its range are *dates*
3. the *"05-05-2021"* literal should be consider a date and not a plain *string*
4. based on the used IRIs, the infection was probably caused by SaRS-CoV-2

RDF provides a formal method to automatically obtain the above statements using the meaning of each of the resources. At the same time, it provides a way to *express* the meaning of resources, again using RDF syntax. As a result, it is the IRIs that appear in the data that determine semantics of a given RDF graph.

The role of IRIs and namespaces

Before going over to the topic of semantics in RDF it is necessary to reiterate over the role of IRIs and namespaces in RDF. Definition 3.2.3 provides the notion of *RDF namespaces*. They play an organizational role in the RDF ecosystem. They are used to group related IRIs and can be treated as *vocabularies of predefined annotations* (refer to Section 3.1.2).

Definition 3.2.3 (RDF namespaces) *An RDF namespace is an abstract container of related IRIs. All IRIs contained within a given IRI share the same prefix, which also identifies the namespace. The party that declared IRIs (resources) in the given namespace is called the maintainer of the namespace.*

Examples of namespaces include:

1. <https://bioportal.bioontology.org/ontologies/ICD10CM/> groups all IRIs related to the ICD-10 CM classification [36]; it thus provides the user with annotations related to *diseases*
2. <http://xmlns.com/foaf/0.1/> groups all IRIs related to the FOAF vocabulary [38] – it contains annotations related to concepts of *persons*

There are three important features of IRIs in the context of RDF semantics:

1. *IRIs* are used to globally identify *resources* and usually take form of a browsable HTTP address
2. anybody can reserve a *namespace* and freely declare concepts (IRIs) in it (as long as it is the namespace is not taken, of course)
3. namespaces and IRIs aimed to be reused by others (such as *vocabularies of concepts*) should be made freely accessible to anybody – e.g. by humans using browsers (in which case an HTML representation of that namespace or IRI should be returned) or by machines (in which case an RDF representation of that namespace or IRI should be returned).

IRIs and RDF namespaces provide foundations for RDF semantics in the following way: the maintainers of the given RDF namespace assign the meaning of the IRIs that belong to the namespace. As mentioned in SubSection 3.2.4., RDF is able to capture meaning of data through annotations using IRIs.

3.2.5. Semantics

The formal semantics provided by the authors of RDF was expressed using *model theory* [41], though presenting it fully is out of scope of this thesis. Nonetheless, we will provide here its practical implications. We will treat the *semantics* of an RDF graph as equivalent to *the set of assertions that can be derived* from it using logical inferences [58].

As argued in Section 3.2.4, the meaning of a given RDF graph emerges from the IRIs that are present in it. The IRIs determine what logical conclusions can be made on the asserted triples. The meaning of the IRIs in turn is provided by the maintainers of namespaces they belong to. Equivalently, their meaning is determined by the list of conclusions that can be carried out if an appropriate set of assertions have appeared in the data.

The given RDF graph can be then transformed using those logical inferences in order to obtain a semantically equivalent graph [41]. Those transformations are also called *inferences* and are the core of semantics of RDF. A specialized class of software capable of conducting automatic entailments on RDF data is called *reasoners*.

RDF semantics in action

Listing 3.6 provides a driving example of how RDF semantics works in practice. At the same time, it presents a typical use case of RDF: expressing data models. The example involves expressing constraints on a hypothetical relationship which models *infection dates of a disease* (denoted *infectionDate* in the example). That relationship has already appeared e.g. in Figure 3.2.

```
@prefix ns:    <http://<url>#> .
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
@prefix foaf:  <http://xmlns.com/foaf/0.1/>
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#>

ns:covidInfectionDate rdfs:domain foaf:Person ;
                      rdfs:range  xsd:date .
```

Listing 3.6: Expressing constraints on the *infectionDate* relationship. It is asserted that its *domain* are *Persons* in the sense of the FOAF vocabulary and that its *range* are XML dates. Both *range* and *domain* concepts were taken from the RDFS vocabulary.

The *infectionDate* relationship from Listing 3.6 is subject to constraints on its *domain* and *range* in the following way. Its *domain* is limited to some definition of a *person* (here – from the FOAF vocabulary [38], more on which will follow later) and its range is limited to *dates* (e.g. as they are defined in the XML Schema Definition [59]). The concepts of the *domain* and *range* of a relationship are in turn provided in the *RDFS* namespace [11]. As a result, any assertions that would state that e.g. a *person* is connected to an *integer* via the *ns:infectionDate* relationship (and not to a *date*) would be considered as data *inconsistency* (*invalid*) under the collective semantics of all IRIs that were involved in the definition of the *infectionDate* relationship.

The example data uses four different namespaces: one to identify the *infectionDate* relationship and three external: *RDFS* [11], *FOAF* [38] and *XSD* [59]. The external namespaces provide definitions of their concepts in RDF themselves and thus can be automatically interpreted by RDF reasoners.

Summary

A usual use case of RDF semantics involves appending the inferred triples to the knowledge base in an iterative manner until no new triples can be added. The triples are generated using the semantics of the IRIs that were present in the dataset. The obtained graph is said to be *inferred* from the initial graph.

The purpose of RDF inferences (and thus of RDF semantics as pointed out in the introduction to Section 3.2.5.) is to make knowledge "hidden" in the graph explicit. To be more precise, the inferred knowledge does not constitute *qualitatively new knowledge* but rather it is a tautologically equivalent to the knowledge already contained within the given RDF graph. Nonetheless, those tautological equivalences help to tackle three problems in data processing in general:

1. model consistency checks – the model can be validated using a reasoner such as Pellet [60]
2. data validation – inference of two contrary assertions can be detected and automatically explained to the user by an RDF reasoner, again using e.g. Pellet [60]
3. simplification of queries – queries on the data can make use of the inferred knowledge

The following list sums up the semantics of RDF in an informal way:

1. *resources* are abstract concepts that are used in domain modeling and their meaning is ceded to their creator; in particular, the creator of the resource defines its semantics, including its associated logical inferences
2. all *literal values* represent database-like values: numbers, dates, strings, XML literals etc.
3. all *blank nodes* and *Internationalized Resource Identifiers* identify resources; the former are used to identify resources solely within the context of a given RDF graph, whereas the latter identify resources globally (on the Web)
4. information can be deduced from an existing *RDF graph* by applying inference rules and thus materializing implicit information within the graph
5. the knowledge materialized through logical inferences can be queried in the same way as the initially asserted data
6. finally, inconsistencies in the data can be detected by verifying the logical consistency of the graph

3.3. The semantic stack of Linked Data

The following section briefly discusses the "*semantic stack*" – i.e. the *various modeling notations and tools, targeting different levels of modeling, available in its ecosystem*. The stack consists of "layers", which are in fact groups of RDF namespaces that handle different levels of abstraction in modeling.

Hendler [61] pointed out that the full description of the semantic stack is unnecessarily complex and needs to be simplified. According to him, it is easy to get lost in the plethora of the available RDF namespaces, its semantic formalisms and technicalities. The following section presents a distilled list of its core concepts.

3.3.1. The four layers

Figure 3.3. presents the order of the layers. They include:

- Layer 0: RDF as the underlying notation (refer to Section 3.2.
- Layer 1: metamodels: RDF Schema [11] and Web Ontology Language [12]
- Layer 2: domain-specific ontologies: ICD-10CM [36], ATC [37] and FOAF [38]
- Layer 3: custom ontologies, used to integrate the data

RDF can be treated as foundations for the rest of layers. Going upwards the stack, OWL and RDFS are *metamodels* that provide the user with model specification capabilities. Domain-specific ontologies in turn use the metadmodels to formalize knowledge in various domains of interest. The integrational layer builds on top of the other three.

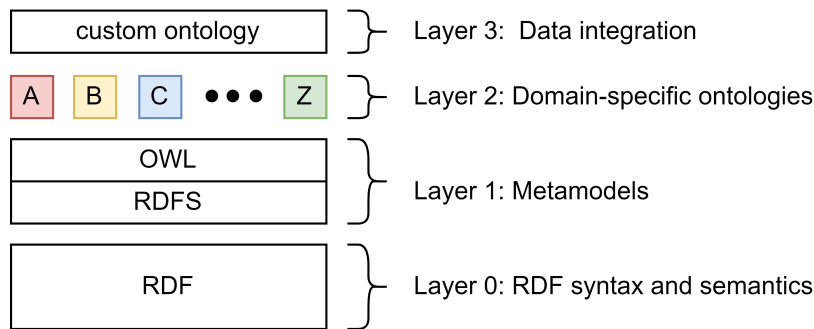


Figure 3.3: The semantic stack of Linked Data.

3.3.2. Layer 0 – RDF as the basis

RDF provides the minimal set of syntactical and semantic rules to annotate the data. For details on RDF, refer to Section 3.2.

3.3.3. Layer 1 – Metamodels: RDFS & OWL

The role of the *metamodels* is to provide a set of annotations (IRIs) that let the user *express models of their data*. The two most widely used metamodels available in the RDF ecosystem are *RDF Schema (RDFS)* and *Web Ontology Language (OWL)*. Horrocks et al. [62] provided a good overview and comparison between the two.

RDFS

RDF Schema (RDFS) [11] is a namespace within the RDF ecosystem providing basic meta-modeling capabilities. Two interesting types of annotations (IRIs) provided by that namespace include:

1. IRIs related to rigorous definition of relationships (including the concepts of *domain* and *range* which have already appeared e.g. in Listing 3.6.)
2. the IRIs related to modeling of hierarchies, e.g. *subClassOf*

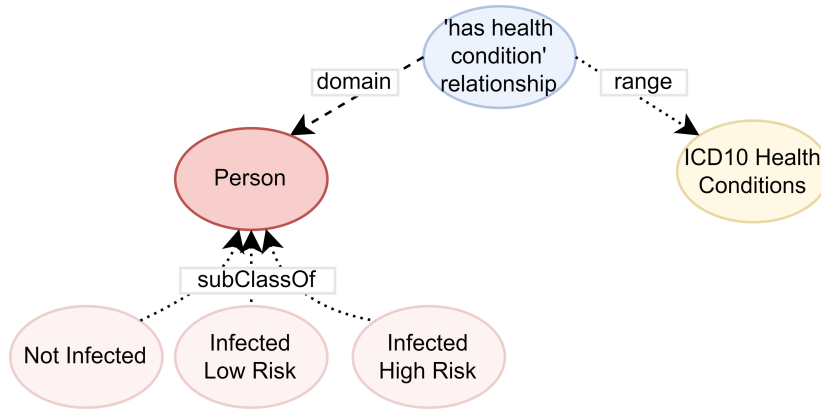


Figure 3.4: Expressivity of RDFS. The example captures its type hierarchy modeling capabilities as well as the ability to define *domains* and *ranges* of relationships.

Figure 3.4. presents an example of application of RDFS to domain modeling. It includes a sample type hierarchy along with a *domain* & *range* specification of a relation.² The presented hierarchy can be practically used in an epidemiological setting to annotate *persons* based on their exposure to the SARS-CoV-2 virus.

Web Ontology Language

Web Ontology Language (OWL) [12] can be regarded as an extension of RDFS which adds more semantic expressivity. It provides the users with a rich vocabulary for defining equivalences between classes and enriching type hierarchies [63]. OWL heavily relies on mathematical formalisms, primarily *description logics* [12] [63] [18].

OWL is suitable for domain modeling which involves classification based on the asserted data. In the context of epidemiology, it enables e.g. formally expressing the following three concepts:

1. "high-risk patients" as those persons "who are COVID-positive AND are desaturated"
2. "low-risk patients" as those persons "who are COVID-positive AND are not desaturated"
3. "not infected patients" as those persons "who are not COVID-positive"

Figure 3.5. provides an example of how inferences within OWL are being made. The provided data has a similar layout and meaning to the example presented in Table 3.1. It adds the two following columns: *symp1* and *symp2*, both storing information about the observed symptoms. Based on the simple domain model of COVID-19 patients defined above and on the asserted list of health conditions and symptoms, the following inferences about the patients can be made using OWL:

1. N9X2 patient belongs to "Infected, High Risk" class
2. A4G2 patient belongs to "Infected, Low Risk" class
3. P0Z2 patient belongs to "Not Infected" class

²The full IRIs of the concepts used in Figure 3.4 include <http://www.w3.org/1999/02/22-rdf-syntax-ns#subClassOf>, <http://www.w3.org/1999/02/22-rdf-syntax-ns#domain> and <http://www.w3.org/1999/02/22-rdf-syntax-ns#range>. For further information on the IRIs provided by RDFS, refer to <https://www.w3.org/TR/rdf-schema/> [11].

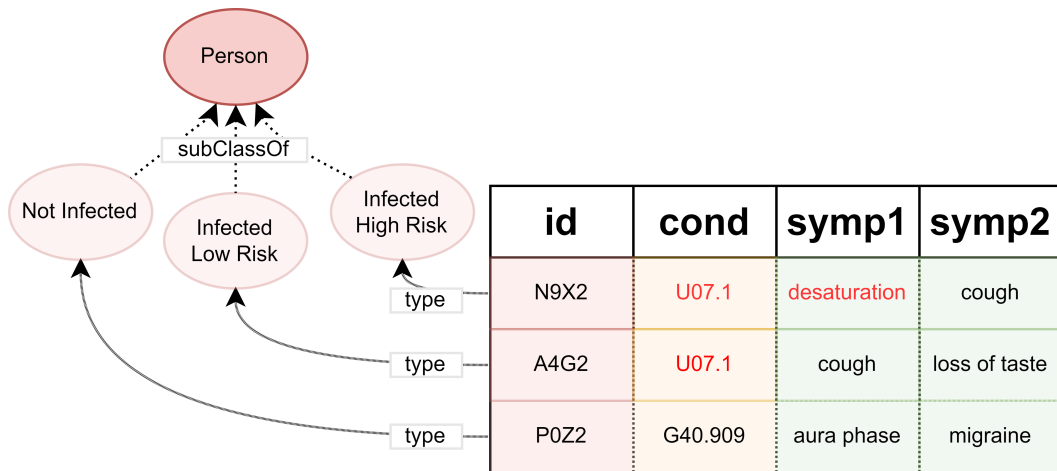


Figure 3.5: Expressivity of OWL. The following inferences about the persons can be made based on the asserted data: N9X2 – "Infected, High Risk" case, A4G2 – "Infected, Low Risk" case, P0Z2 – "Not Infected" case. Red font denotes premises of belonging to the "Infected, High Risk" class. The "U07.1" stands for Diagnosed COVID-19 in the ICD-10CM taxonomy [36].

3.3.4. Layer 2 – Domain-specific ontologies

The role of *domain-specific ontologies* is to provide formalized vocabularies for various *domains* – e.g. biomedicine, sociology, chemistry etc.

Ontologies

According to Gruber [35], ontologies are "*specifications of a conceptualization*". To put it simply, they provide formalized vocabularies of terms with their strictly defined meaning. They aim at formalizing knowledge in order to avoid logical errors and inconsistencies when modeling phenomena. Uschold and Grüninger [64] provided a good overview of principles, methods and applications of ontologies.

The following terms are all synonyms for *ontology*, used in various contexts: *taxonomy*, *hierarchy*, *domain model* [35]. Moreover, ontologies have a special relationship to *database schemas*. Uschold [49] provided a good overview of that topic. According to him, both share many common features (e.g. have strong foundations in formal logic) but differ in one key aspect: ontologies are focused on the *meaning* whereas database schemas are focused on *data*.

Domain-specific ontologies

Domain-specific ontologies are ontologies that concentrate on a single, specific domain of interest. They play the role of controlled vocabularies. The following list presents examples of domain-specific ontologies used in the context of the epidemiological study at the Pediatric Hospital of the Medical University of Warsaw:

1. Friend-of-a-Friend (FOAF) [38] – a vocabulary of concepts related to *persons*, including attributes like *age* and *gender* and some social relationships, like the '*knows*' relationship
2. Anatomical Therapeutical Chemical Classification [37] – a vocabulary of chemical substances used in medicine

3. International Classification of Diseases, v10, Clinical Modification (ICD-10CM) [36] – a vocabulary of human diseases

3.3.5. Layer 3 – Data integration

Lenzerini [65] defines data integration as "*the problem of combining data residing at different sources and providing the user with a unified view of these data*". In the context of Linked Data, it practically means to combine two domain-specific ontologies in order to design a novel, interdisciplinary model of some phenomena. This layer covers conceptually everything in the data model under design that requires using at least two different domain-specific ontologies. It boils down to defining relationships between concepts from two different ontologies.

Chapter 4

Results

In the following chapter, results of the thesis regarding design and implementation of the semantic data model for the epidemiological study follow. Additionally, a statistical verification of the possible risk factors on COVID-19 severity are provided.

4.1. Data model: the *covidepid* ontology

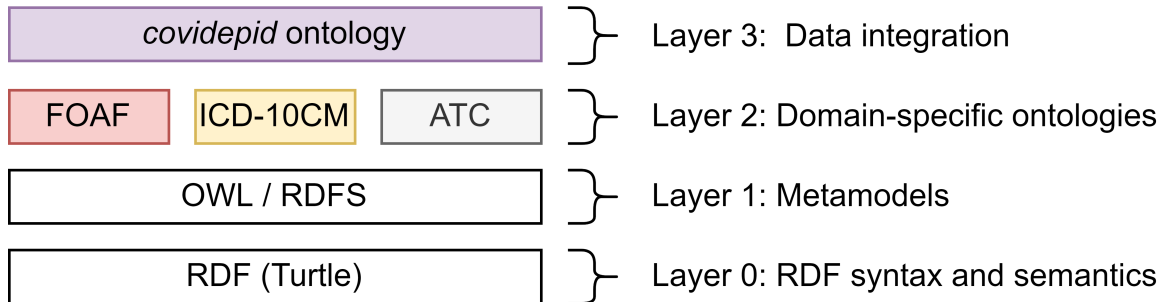


Figure 4.1: Layers of the obtained data model.

As pointed out in Section 1.3., one of the goals of the thesis was to create a data model linking three domain-specific ontologies: ICD-10CM [36], ATC [37] and FOAF [38]. Figure 4.1. presents the layers of the obtained model (refer to Section 3.3. for more details on the role of different semantic layers in Linked Data). The *covidepid* ontology¹ merged the other three ontologies using RDFS [11]. Turtle has been chosen as the target concrete RDF syntax (refer to Section 3.2.3.) to store the obtained RDF data.

Figure 4.2. presents the Linked Data model used in the epidemiological study. To improve readability, a UML-like [66] notation has been used. The rectangles represent concepts. Entries on the white background within rectangles represent the *literal values* associated with the concept (refer to Section 3.2.2. for the notion of literal values). Arcs between rectangles represent possible relationships between instances of the concepts they connect. Different colors denote different ontologies.

¹<https://github.com/kubajal/covidepid/releases/download/v0.1/covidepid.owl>

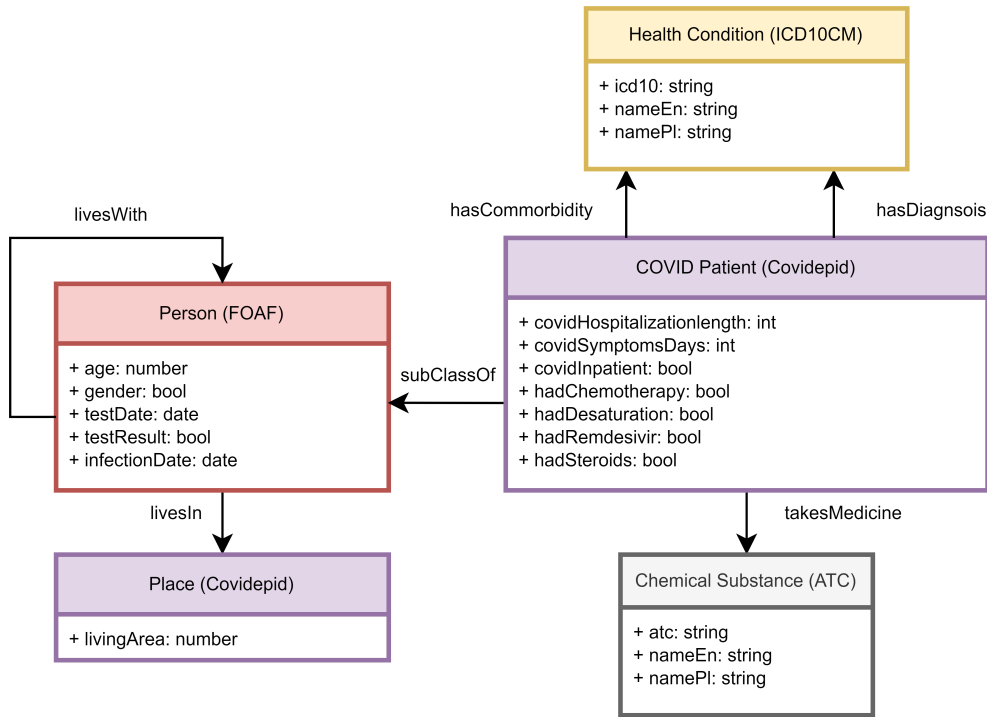


Figure 4.2: High-level overview of the data model used in the study. The concepts of *Disease*, *Medicine* and *Person* were taken from the ICD-10CM [36], ATC [37] and FOAF [38] domain-specific ontologies. The custom *covidepid* ontology plays an integrational role.

4.2. Data pipeline

Table 4.1. recaps the elements of the implemented data pipeline, presented already in the methodological part of the thesis in Figure 2.1 (refer to Section 4.2.). A more detailed description of each step follows.

Step 1: Data collection through a user interface

React JSON Schema Form² has been used as a rapid application development framework to implement the user interface for entering the survey data. The features of the user interface included:

1. entering a new record
2. editing an existing record
3. listing all existing records by their *identifiers*, *creation date* or *creator*

One of the main challenges for this step was to detect possible collisions in the case when two parties would try to edit the same record in parallel. The problem was solved by checking the MD5 hash used as a basis of their edition.

As a result of this step, a JSON file was produced for each record entered through the user interface (see Step 1. in Table 4.1. and in Figure 2.1).

²<https://github.com/rjsf-team/react-jsonschema-form>

step number	method	input	result
1.	user interface	paper surveys	JSON data
2.	a Python program	JSON data	RDF data
3.	Apache Jena	RDF data	the RDF data extended by <i>RDF inferences</i>
4.	RStudio	the RDF data extended by <i>RDF inferences</i>	insights about the data

Table 4.1: Partial results obtained in each step of the data pipeline. Refer to Figure 2.1 as a reference architecture of the system.

Step 2: JSON-to-RDF mapping

Consequently, the obtained JSON files were transformed into a single RDF file using the *rdflib* v6.0.1 [67] library in Python.

The library provides a number of methods for handling RDF data. It has built-in support for various RDF namespaces, including *RDFS* [11] and *OWL* [12]. Additionally, it provides export to a few data serialization formats: *Turtle* [10] and *RDF/XML* [54] among others. It also provides limited support for SPARQL queries and RDF entailments [48]. It is suitable for simple processing and extraction of RDF data. It was used in the thesis primarily as a method of transformation between JSON and RDF representation of the epidemiological data.

As a result of this step, all JSON files obtained in the previous step were transformed into a single RDF file (see Step 2. in Table 4.1. and in Figure 2.1).

Step 3: Loading the RDF data into a triple store

The next step involved loading the RDF file into a triple store (an RDF database) called Apache Jena [45] which has been configured to conduct inferences on imported data using an RDFS reasoner³⁴. The Apache Jena triple store was exposed as a REST API using the Fuseki wrapper⁵. This enabled seamless integration of Apache Jena with external software, such as RStudio.

This step provided inferences, primarily for the observed comorbidities and medicines (see Step 3. in Table 4.1. and in Figure 2.1). It made it distinctly easier to answer the following questions later on:

1. how many patients had *immunological* diseases?
2. how many patients were treated continually using *cardiovascular system drugs*, unrelated to COVID-19?

³<https://jena.apache.org/documentation/inference/#rdfs>

⁴The RDFS Reasoner IRI: <http://jena.hpl.hp.com/2003/RDFSExptRuleReasoner>

⁵<https://jena.apache.org/documentation/fuseki2/fuseki-configuration.html>

Step 4: Data analysis

Finally, RStudio was used to query Apache Jena using SPARQL [48] and statistically analyze the retrieved data using the *coin* package [68] (see Step 4. in Table 4.1. and in Figure 2.1). Refer to Section 2.4. for more information on the methods of hypothesis testing used in the thesis.

4.3. Risk factors of severe COVID-19 in children

In the following section, an analysis of the influence of patient's comorbidities on COVID-19 severity was verified using statistical tests. The input of the statistical analysis consisted of data provided by the Linked Data pipeline, as was implemented in Section 4.2.

4.3.1. High-level overview of the data

The data collected in the first phase of the epidemiological study numbered 303 cases of COVID-positive inpatients that were admitted to the Pediatric Hospital of the Medical University of Warsaw between December 2020 and July 2021. Figure 4.3. depicts the partitioning of the sample subject to the thesis. Of the 303 patients, 90 fulfilled the requirements to be considered in the COVID-19 severity analysis. The requirements included:

1. COVID-19 as the main diagnosis
2. passing the preselection in the Emergency Room – the patient required hospitalization and was admitted to the hospital

Figure 4.4. presents the histogram of patients' hospitalization lengths. The distribution fails the Shapiro-Wilk test of normality available in the R package [46] with the p-value < 0.01 .

Figure 4.5 presents the observed counts of comorbidities per disease class, according to the ICD-10CM ontology [36]. All comorbidity classes except for two turned out to be under-represented in the observed sample to be taken into account as a possible risk factor of COVID-19 severity in children (refer to the methodology of the statistical tests in Section 2.4).

4.3.2. Hypothesis testing

Based on the obtained comorbidity class counts, only the following two classes passed the minimal criteria regarding the observed counts: E00-E89 – *Endocrine, nutritional and metabolic diseases* and Q00-Q99 – *Congenital malformations, deformations and chromosomal abnormalities*.

Table 4.2. presents the results of the Mann-Whitney-Wilcoxon U test for both investigated (*risk factor - hospitalization length*) pairs. The tests were conducted with the significance level set to 0.05.

The results indicate that patients E00-E89 – *Endocrine, nutritional and metabolic diseases* had a statistically different length of hospitalization compared to patients that did not have this class of comorbidities.

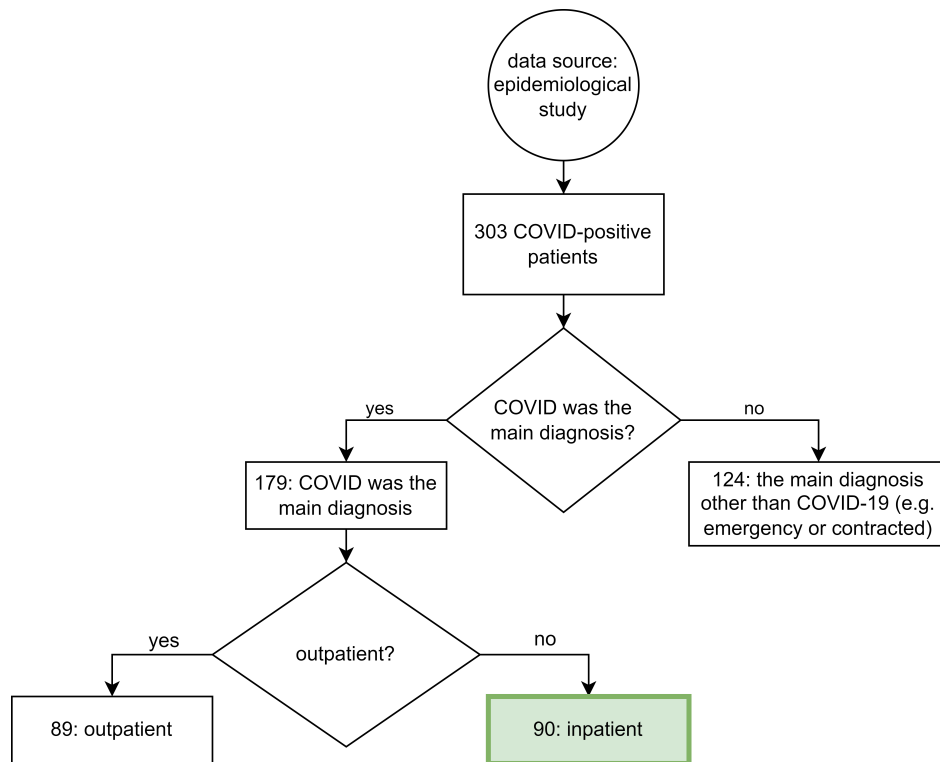


Figure 4.3: Classification of patients. The sample consisted of 90 patients registered in the epidemiological study that 1) had COVID-19 as their main diagnosis 2) were severe enough to be admitted to the Pediatric Hospital. The data were obtained between December 2020 and July 2021. The sample investigated in the thesis is marked green.

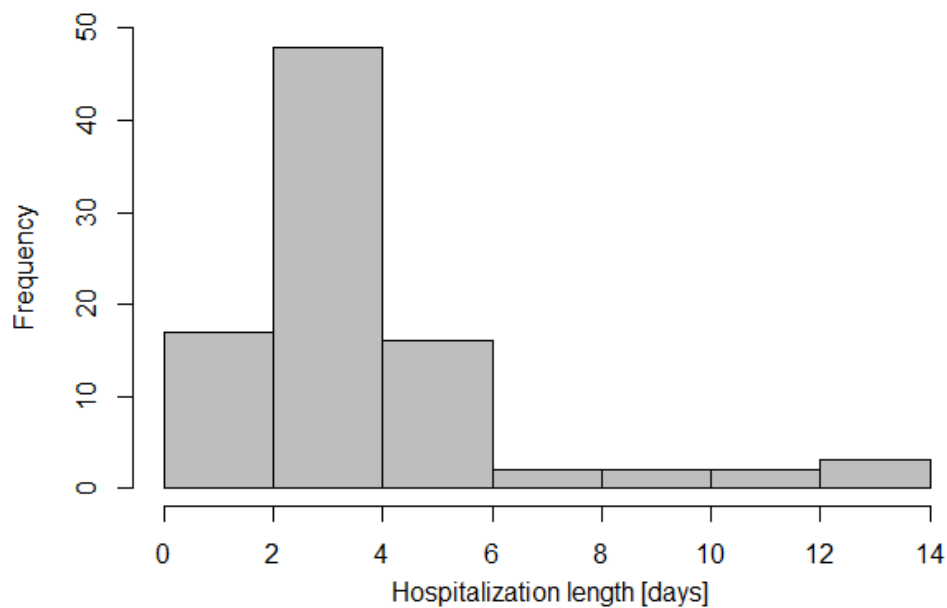


Figure 4.4: The histogram of hospitalization lengths. The p-value obtained using Shapiro-Wilk's normality test was smaller than 10^{-10} .

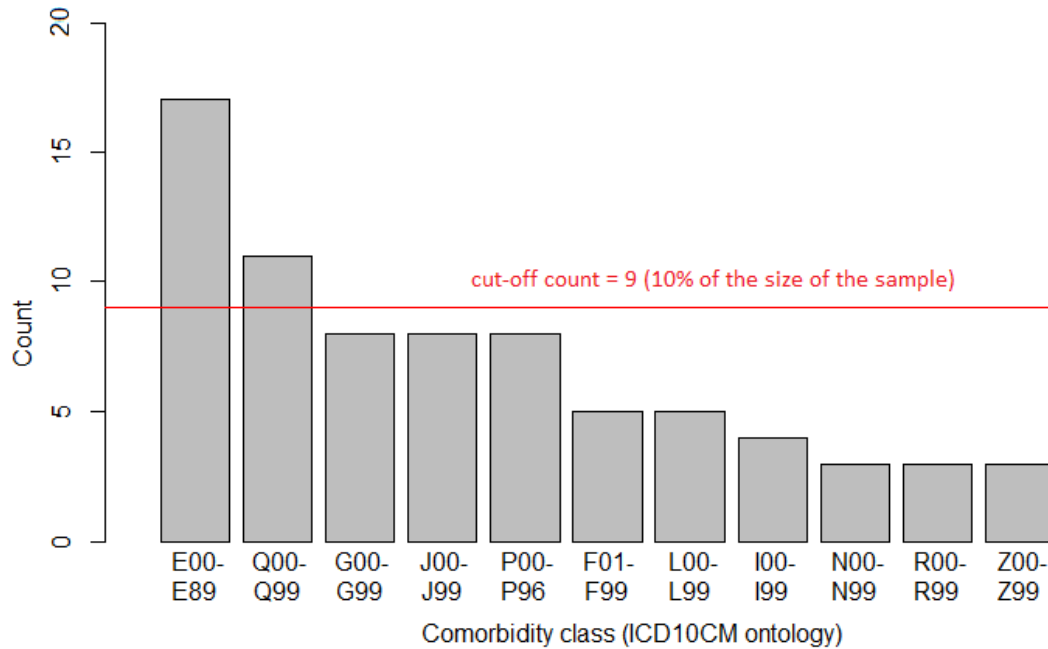


Figure 4.5: Counts of the observed comorbidity classes in the study sample of 90 patients. Only two pass the required minimum count of observations equal to 9 (refer to Section 2.4.): E00-E89 – *Endocrine, nutritional and metabolic diseases* and Q00-Q99 – *Congenital malformations, deformations and chromosomal abnormalities*

	Q00-Q99 <i>Congenital malformations, deformations and chromosomal abnormalities</i>	E00-E89 <i>Endocrine, nutritional and metabolic diseases</i>
Hospitalization length	0.17	0.005

Table 4.2: Results of the Mann-Whitney-Wilcoxon U test of hospitalization lengths in two subpopulations. Table cells contain the reported p-values. Cells with results considered statistically significant were marked green, otherwise the cell was marked red. The significance level was set to 0.05.

Chapter 5

Discussion

This chapter evaluates the usefulness of Linked Data as a data-processing paradigm. It reviews its features in the context of:

1. its data-processing capabilities, including the results obtained in Section 4.3.
2. the design (Section 4.1.) and implementation (Section 4.2.) of the obtained data model

5.1. The role of Linked Data in the epidemiological study

The most interesting use case of the Linked Data approach during the analysis of COVID-19 severity in children was the calculation of the number of comorbidities per disease class. These kinds of multi-domain, analytical queries is where this approach to data processing shines. In the case of this study, it enabled cross-sections of the data in a presentable way. The results of the query are presented in Figure 4.5.

Figure 5.1 depicts the problem of counting the comorbidities *per disease class*. It involved:

1. integrating the anonymized raw data about patients' comorbidities, provided by the Medical University of Warsaw, with the ICD-10CM taxonomy of diseases
2. querying the asserted data and its hierarchy of classes (the ICD-10CM taxonomy) at the same time in order to count the comorbidities in a *per disease class* manner

The asserted relationships are marked with solid lines. The dashed line represents inferences that can be made about *disease B* that distinctly simplifies the aggregation in the following way. Without the inference, the aggregation would have to be formulated separately *for each possible height of the hierarchy subtree*. On the other hand, due to the transitive nature of the 'is subclass of' relationship, a direct link between *disease B* and the hierarchy root (*E00-E89*) can easily be inferred. The link can then be reused as an invariant of a single query, expressed in SPARQL [48], tremendously simplifying data processing.

Listing 5.1. presents the SPARQL query, used to obtain statistics of comorbidities. In order to improve readability, Line 1. substitutes the list of prefixes for namespaces of RDF Schema [11] and the *covidepid* ontology. The 3. line returns the counts of comorbidities, aggregated per ICD-10CM diseases. Lines 5-6 match any patients with the classes of their comorbidities. Lines 8. and 9. constrain the results to only those disease classes, which do not have a direct antecedent of degree 2 in the hierarchy of disease classes (i.e. they are on the first level of the hierarchy class). ¹.

¹A technical note for advanced users of RDFS: the query in Listing 5.1 omits a few technical clauses in

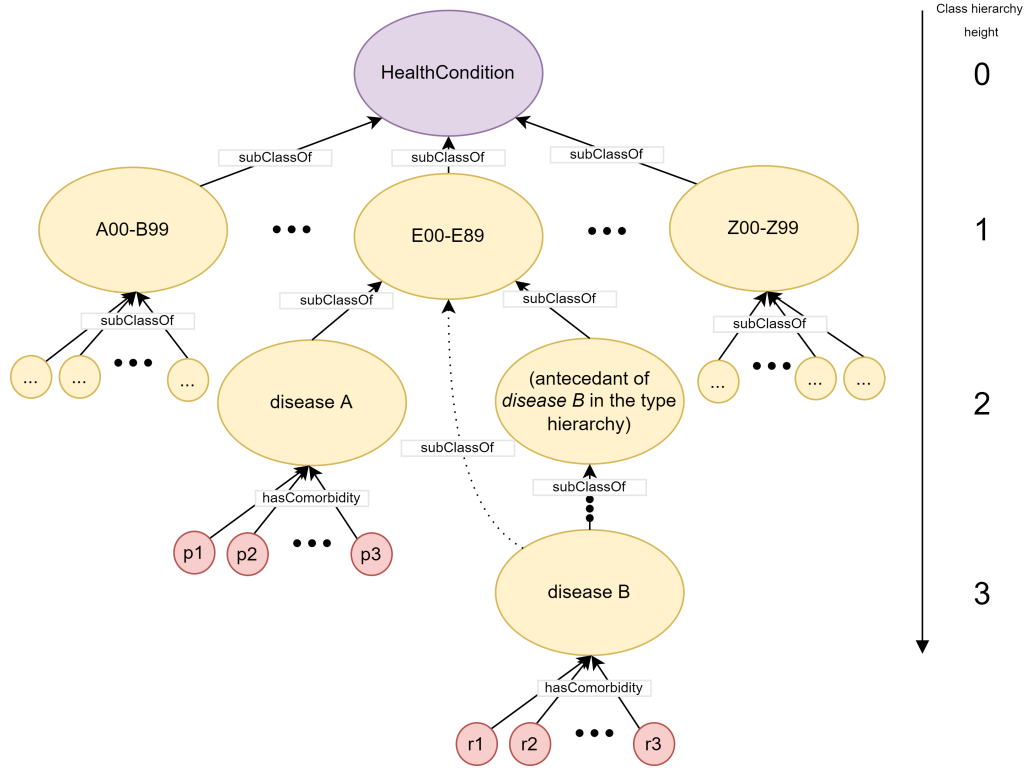


Figure 5.1: Varying depth of class hierarchy is the problem of calculating counts of the *hasComorbidity* relationship for the E00-E89 – *Endocrine, nutritional and metabolic diseases* class of diseases. The asserted data are represented using solid lines. Without the inferences (the dashed line), different formulations of the aggregating query are needed for all subtree heights. Colors have the following meaning – yellow: ICD-10CM disease classes, red: FOAF ontology (patients), purple: the integrational layer (the *covidepid* concepts).

Listing 5.1: Aggregation of comorbidities in SPARQL [48]. Calculation of comorbidity class frequency using class hierarchy traversal along the triples inferred using transitivity of the *rdfs:subClassOf* relation.

```

1 <... prefixes ... >
2
3 SELECT ?healthConditionClass (count(?healthConditionClass) as ?cnt)
4 WHERE {
5     ?patient covidepid:hasComorbidity ?disease .
6     ?disease rdfs:subClassOf ?diseaseClass .
7     FILTER NOT EXISTS {
8         ?diseaseClass rdfs:subClassOf ?diseaseClassParent .
9         ?diseaseClassParent rdfs:subClassOf covidepid:HealthCondition .
10    }
11 }
12 GROUP BY ?diseaseClass

```

the query that filter out: 1) *rdfs:Resource* as the root of the inheritance hierarchy, 2) reflexivity of the *rdfs:subClassOf* relation.

5.2. Statistical results

The results of a statistical analysis of the obtained data (Section 4.3.) suggests that there is a significant difference between hospitalization lengths of patients with the E00-E89 – *Endocrine, nutritional and metabolic diseases* compared to those without such a comorbidity ($p\text{-value} = 0.005$).

The obtained statistical result indicates the fundamental role Linked Data could play in epidemiology. It should not be treated however as a final conclusion about COVID-19 severity. The reasons are three-fold:

1. the level of granularity when counting disease classes is very low; it treats diseases of various etiology within the same subtree in the same way
2. the study sample size (90) was small in comparison to the related work ([27], [29], [25])
3. patient’s hospitalization length is a very vague indicator of COVID-19 severity

Regarding the first point, other aggregation levels of the investigated comorbidities in the ‘subclass of’ hierarchy are also possible, e.g. the level 2 or 3. This would require a bigger study sample though as the sample obtained in the first phase of the study is quite small (90 cases) and there are currently not enough cases to aggregate reasonably on other hierarchy levels. Provided there is more data available, it is trivial to reformulate the query in Listing 5.1. to also capture higher levels of granularity.

Regarding the second point, further monitoring of the patients admitted to the hospital is desired in order to increase the studied sample. With the ongoing second phase of the epidemiological study, new cases will be provided shortly. The data pipeline implemented in Section 4.2. will facilitate curation of their data.

Regarding the third point, the investigated COVID-19 severity indicators could also include other attributes of patient’s hospitalization, such as: *number of symptoms, occurrence of desaturation, blood test results*, among others. As a rule of thumb, the more relationships would turn out to be statistically significant for a given comorbidity class, the stronger premises the comorbidity class would have to be a risk factor. In the case of blood tests, integration of new types of data would be required into the already existing model. With the gained experience regarding modeling using the Linked Data approach in the *covidepid* ontology, this seems to be feasible.

5.3. Technical aspects

The bulk part of the work in the thesis was invested into implementation of data collection and its curation as described in Section 2.1. It required a lot of time and dedication to:

1. align two different programs to use the same data conceptual model: the user interface on one hand and the JSON-to-RDF transformer on the other hand
2. mix four different kinds of software in the data pipeline: a JavaScript program, a Python program, Apache Jena and RStudio

As a result, the implemented data transformations should be radically simplified. This can be achieved by using software such as the proprietary REDCap² or OpenClinica³ (GNU

²<https://www.project-redcap.org/>

³<https://github.com/OpenClinica/OpenClinica>

LGPL license). It remains unclear how those programs could provide export of the stored data into an RDF format, e.g. Turtle [10].

Alternatively, a different approach to data transformations could be taken. Instead of mapping the JSON files to RDF using a custom Python program written in the *rdflib* v6.0.1 [67] library, a set of *RML*⁴ rules for data transformations could be defined. The *RDF Mapping Language* (*RML*) is a generic mapping language, aiming at expressing transformation rules of various non-RDF data formats to RDF. The entry threshold of this approach seems low in comparison with the custom-code-based solution for JSON-to-RDF data transformation. Unfortunately, using RML assumes using the custom JavaScript-based user interface that produces the JSON files at the same time, whereas an all-in-one solution would be desired.

Summing up, the Linked Data approach requires a lot of effort in the technical aspect of data curation. In the end, the user is often left with a multi-staged data pipeline that is complex to maintain. Although this thesis provides a proof of concept of such a Linked Data curation pipeline, further research has to be done in the area of tools aimed specifically at epidemiology.

⁴<https://rml.io/docs/>

Chapter 6

Conclusions

This chapter sums up the results obtained in the thesis and concludes them. It emphasizes the crucial role that the Linked Data approach played in obtaining them.

6.1. Linked Data in epidemiology

Overall, the two main features of Linked Data as a tool in data management are:

1. the syntax for both metadata and data through the annotations using Internationalized Resource Identifiers
2. the semantic data model based on graph theory and computational logics

Those features enable the following advantages in data processing over the traditional semantic-agnostic approach to data management:

1. seamless integration of various domain-specific ontologies that provide meaning to the data
2. relying on a declarative data model rather than on iterative data processing in order to extract cross-sections from the data

The thesis presented a minimal example of benefits that Linked Data could provide in the context of an epidemiological study. In the case of the study at the Pediatric Hospital of the Medical University of Warsaw, the Linked Data paradigm enabled integration of various domain-specific ontologies by providing a common standard to express links between heterogeneous data. It was used as a framework for the integration of various domain-specific ontologies, including: ICD-10CM [36] and FOAF [38].

Additionally, by bringing machine-interpretable semantics to the data, the Linked Data approach enabled answering an analytical query concerning counting comorbidities in COVID-positive patients in a *per ICD-10CM disease class* manner. Using semantics of the data in an explicit way, it was easy to formulate an aggregating cross-section of the asserted data that involved traversing complex hierarchy classes. The query used knowledge inferred from the asserted data. The obtained query result was used to assess the influence of the comorbidity type on the patients' hospitalization length. It transpired that there is a statistically significant difference between the median of hospitalization length in patients with metabolic diseases compared to those that do not have such comorbidities.

Ultimately, Linked Data enables concise formulation of queries that would otherwise require multi-staged data processing. As a result, introducing formal semantics to the data and means to integrate various domain-specific ontologies the way that the Linked Data paradigm does, can be viewed as a massive improvement to data analytics in general.

On the other hand, some knowledge of mathematical formalisms, such as graph theory, description logics and computational complexity is advised when dealing with Linked Data as the paradigm heavily relies on those mathematical formalisms. It has a steep learning curve which entry-level users may find discouraging. The user can quickly get overwhelmed by the syntax of IRIs and RDF, nitty-gritty details of formal specifications and analysis of time complexity of the algorithms performed under the hood when conducting semantic inferences. Additionally, there is a lack of high-quality open-source user interfaces that would enable easy collection and curation of epidemiological surveys in a Linked Data format. Further efforts should be put into solving that problem. Finally, the current state of the open-source modeling tools available in the Linked Data community leaves a lot to be desired.

Nonetheless, RDF triples stores – such as Apache Jena [45] – together with reasoners – such as Pellet [60] – provide an exciting alternative to traditional data stores. They enable seamless integration of both data and metadata from different domains. Thus, the Linked Data paradigm may find users primarily in an interdisciplinary setting, including epidemiology.

6.2. Future work

In order to provide high-quality results of the epidemiological study at the Pediatric Hospital of the Medical University of Warsaw, future work requires extending COVID-19 severity indicators definition by additional attributes of patient’s hospitalization, such as blood tests, occurrence of desaturation, etc. Additionally, the second phase of the epidemiological study ought to be finished in the upcoming months. It will extend the already available study sample, enabling more confident statistical results.

The future work will entail:

1. verifying the impact of continuous use of various medicine types on the COVID-19 outcomes in children patients
2. investigating COVID-19 spread dynamics based on the timeline of infections within the patient’s family

The results of the epidemiological study will be published in a scientific article.

List of Figures

1.1. Phases of the epidemiological study	8
2.1. Data processing pipeline	13
3.1. An example of an <i>RDF triple</i> represented as a graph	18
3.2. An example of an RDF graph	19
3.3. The semantic stack of Linked Data	27
3.4. Expressivity of RDFS	28
3.5. Expressivity of OWL	29
4.1. Layers of the obtained data model	31
4.2. High-level overview of the data model used in the study	32
4.3. Classification of the study sample cases	35
4.4. Histogram of hospitalization lengths	35
4.5. Summary of the observed comorbidities	36
5.1. Varying depth of the taxonomy of diseases according to ICD-10CM	38

List of Tables

2.1. The scheme of the investigated statistical hypotheses	13
3.1. Raw data vs. data with annotations	17
3.2. A Subset of Extended Backus-Naur Form	22
4.1. Partial results obtained in each step of the data pipeline	33
4.2. Results of the Mann-Whitney-Wilcoxon U test	36

Bibliography

- [1] World Health Organization. Coronavirus disease (COVID-19) – Overview. <https://www.who.int/health-topics/coronavirus>, accessed on: 14th December 2021.
- [2] Avneet Kaur et al. COVID-19 Infection: Epidemiology, Virology, Clinical Features, Diagnosis and Pharmacological Treatment. *Current Pharmaceutical Design*, 27, 01 2021.
- [3] World Health Organization. COVID-19 disease in children and adolescents: Scientific brief. 09 2021.
- [4] World Health Organization. WHO Coronavirus (COVID-19) Dashboard. <https://covid19.who.int>, accessed on: 12th December 2021.
- [5] Maria Nicola et al. The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *Int. J. Surg.*, 78:185–193, June 2020.
- [6] World Wide Web Consortium. What is Linked Data? <https://www.w3.org/standards/semanticweb/data>, accessed on: 14th December 2021.
- [7] Mark Wilkinson et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 03 2016.
- [8] Maulik Kamdar, Javier Fernández, Axel Polleres, Tania Tudorache, and Mark Musen. Enabling Web-scale data integration in biomedicine through Linked Open Data. 2, 09 2019.
- [9] Richard Cyganiak, David Hyland-Wood, and Markus Lanthaler. RDF 1.1 Concepts and Abstract Syntax. *W3C Recommendation*, 02 2014.
- [10] Eric Prud’hommeaux and Gavin Carothers. RDF 1.1 Turtle. *W3C Recommendation*, 02 2014.
- [11] Ramanathan Guha and Dan Brickley. RDF Schema 1.1, 02 2014.
- [12] Sean Bechhofer, Frank Harmelen, James Hendler, Ian Horrocks, Deborah McGuinness, Peter Patel-Schneider, and Lynn Stein. OWL Web Ontology Language Reference. 02 2004.
- [13] Tim Berners-Lee. Linked data - design issues, 2006. <https://www.w3.org/DesignIssues/LinkedData.html>, accessed on: 18th October 2021.
- [14] Tim Berners-Lee. What the Semantic Web Can Represent. 01 1998.

- [15] Tim Berners-Lee and Robert Cailliau. WorldWideWeb - Proposal for a HyperText Project, 1990. <https://www.w3.org/Proposal.html>, accessed on: 19th October 2021.
- [16] T Berners-Lee, James Hendler, and Olli Lassila. The Semantic Web. *Scientific American*, 284:35, 01 2001.
- [17] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data: The Story so Far. *International Journal on Semantic Web and Information Systems*, 5:1–22, 07 2009.
- [18] Dean Allemang and James Hendler. *Semantic web for the working ontologist - modeling in RDF, RDFS and OWL*. 01 2008.
- [19] Aakash Ahmad et al. An Overview of Ontologies and Tool Support for COVID-19 Analytics. 10 2021.
- [20] Leila Bayoudhi, Najla Sassi, and Wassim Jaziri. An Overview of Biomedical Ontologies for Pandemics and Infectious Diseases Representation. *Procedia Computer Science*, 192:4249–4258, 01 2021.
- [21] Biswanath Dutta and Michael Debellis. CODO: An Ontology for Collection and Analysis of Covid-19 Data. 09 2020.
- [22] Schriml Lynn et al. Disease Ontology: A backbone for disease semantic integration. *Nucleic acids research*, 40:D940–6, 11 2011.
- [23] Yongqun He et al. CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. *Scientific Data*, 7, 12 2020.
- [24] Lindsay Kim et al. Hospitalization Rates and Characteristics of Children Aged <18 Years Hospitalized with Laboratory-Confirmed COVID-19 - COVID-NET, 14 States, March 1-July 25, 2020. *MMWR. Morbidity and mortality weekly report*, 69, 08 2020.
- [25] Miranda Delahoy et al. Hospitalizations Associated with COVID-19 Among Children and Adolescents — COVID-NET, 14 States, March 1, 2020–August 14, 2021. *MMWR. Morbidity and Mortality Weekly Report*, 70, 09 2021.
- [26] Centers for Disease Control and Prevention. COVID-NET: COVID-19-Associated Hospitalization Surveillance Network. <https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covid-net/purpose-methods.html>, accessed on: 12th December 2021.
- [27] Micaela Sandoval, Duc T. Nguyen, Farhaan S. Vahidy, and Edward A. Graviss. Risk factors for severity of COVID-19 in hospital patients age 18–29 years. *PLOS ONE*, 16(7):e0255544, July 2021.
- [28] Rebecca Woodruff et al. Risk Factors for Severe COVID-19 in Children. *Pediatrics*, 10 2021.
- [29] Sara Rubenstein et al. COVID-19 in Pediatric Inpatients: A Multi-Center Observational Study of Factors Associated with Negative Short-Term Outcomes. *Children*, 8(11), 2021.
- [30] Kristina Gaietto et al. Asthma as a risk factor for hospitalization in children with COVID-19: A nested case-control study. *Pediatric allergy and immunology*, 11 2021.

- [31] Sara Assaf et al. Asthma and severe acute respiratory syndrome coronavirus 2019: current evidence and knowledge gaps. *Current Opinion in Pulmonary Medicine*, Publish Ahead of Print, 10 2020.
- [32] Krishan Chhibba et al. Prevalence and characterization of asthma in hospitalized and non-hospitalized patients with COVID-19. *Journal of Allergy and Clinical Immunology*, 146, 06 2020.
- [33] Eduardo Garcia-Pachon et al. Asthma prevalence in patients with SARS-CoV-2 virus infection detected by RT-PCR not requiring hospitalization. *Respiratory Medicine*, 171:106084, 07 2020.
- [34] Manon Grandbastien et al. SARS-CoV-2 Pneumonia in Hospitalized Asthmatic Patients Did Not Induce Severe Exacerbation. *The Journal of Allergy and Clinical Immunology: In Practice*, 8, 06 2020.
- [35] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5:199–220, 1993.
- [36] World Health Organization. ICD-10: international statistical classification of diseases and related health problems: tenth revision, 2004.
- [37] WHO Collaborating Centre for Drug Statistics Methodology. ATC classification index with DDDs, 2021., 2020.
- [38] Dan Brickley and Libby Miller. FOAF Vocabulary Specification. Namespace Document 2 Sept 2004, FOAF Project, 2004. <http://xmlns.com/foaf/0.1/>.
- [39] W3C. Extensible Markup Language (XML) 1.0 (Fifth Edition). 01 2008.
- [40] Kenneth Ross and Charles Wright. *Discrete Mathematics*, volume 70. 01 1992.
- [41] Patrick Hayes and Peter Patel-Schneider. RDF 1.1 Semantics. *W3C Recommendation*, 02 2014.
- [42] Mark A. Musen. The protégé project: a look back and a look forward. *AI Matters*, 1(4):4–12, 2015.
- [43] Patricia Whetzel et al. BioPortal: Enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic acids research*, 39:W541–5, 06 2011.
- [44] Dirk Merkel. Docker: lightweight linux containers for consistent development and deployment. *Linux journal*, 2014(239):2, 2014.
- [45] Apache Software Foundation. Apache Jena. <https://jena.apache.org/>, accessed on: 14th December 2021.
- [46] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.
- [47] Pascal Hitzler. A review of the semantic web field. *Communications of the ACM*, 64:76–83, 01 2021.
- [48] Steve Harris and Andy Seaborne. SPARQL 1.1 Query Language, March 2013.

- [49] Michael Uschold. Ontology and database schema: What's the difference? *Applied Ontology*, 10:243–258, 12 2015.
- [50] Stefan Decker, Sergey Melnik, Frank Harmelen, Dieter Fensel, Michel Klein, Michael Erdmann, and Ian Horrocks. The semantic web: the roles of XML and RDF. *IEEE Internet Computing*, 4, 10 2000.
- [51] Cory Doctorow. Metacrap: Putting the torch to seven straw-men of the meta-utopia. <http://www.well.com/doctorow/metacrap.htm>, accessed on: 14th December 2021.
- [52] Aaron Swartz. Aaron Swartz's A Programmable Web: An Unfinished Work. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 3:1–64, 02 2013.
- [53] David Hyland-Wood. What's New in RDF 1.1. *W3C Working Group Note*, 02 2014.
- [54] Fabien Gandon and Guus Schreiber. RDF 1.1 XML Syntax. *W3C Recommendation*, 02 2014.
- [55] Robin J. Wilson. *History of Graph Theory*. CRC Press, 2013.
- [56] Lisa Ehrlinger and Wolfram Wöß. Towards a Definition of Knowledge Graphs. 09 2016.
- [57] Manu Sporny, Dave Longley, Gregg Kellogg, Markus Lanthaler, and Niklas Lindström. JSON-LD 1.1. A JSON-based Serialization for Linked Data. *W3C Recommendation*, 02 2014.
- [58] Drew McDermott and Dejing Dou. Representing Disjunction and Quantifiers in RDF. volume 2342, pages 250–263, 06 2002.
- [59] Shudi Gao, C. M. Sperberg-McQueen, Henry Thompson, Noah Mendelsohn, David Beech, and Murray Maloney. W3C XML Schema Definition Language (XSD) 1.1 Part 1: Structures. *W3C Recommendation*, 04 2012.
- [60] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. Pellet: A practical owl-dl reasoner. *Web Semantics*, 5(2):51–53, June 2007.
- [61] James Hendler. Tonight's Dessert: Semantic Web Layer Cakes. page 1, 05 2009.
- [62] Ian Horrocks, Peter Patel-Schneider, and Frank Harmelen. From SHIQ and RDF to OWL: the making of a web ontology language. *Web Semantics: Science, Services and Agents on the World Wide Web*, 1:7–26, 07 2003.
- [63] Ian Horrocks. Owl: A description logic based ontology language. In Peter van Beek, editor, *Principles and Practice of Constraint Programming - CP 2005*, pages 5–8, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.
- [64] Michael Uschold and Michael Grüninger. Ontologies: Principles, methods and applications. *The Knowledge Engineering Review*, 11, 01 1996.
- [65] Maurizio Lenzerini. Data Integration: A Theoretical Perspective. pages 233–246, 01 2002.
- [66] James Rumbaugh, Ivar Jacobson, and Grady Booch. *Unified Modeling Language Reference Manual, The (2nd Edition)*. Pearson Higher Education, 2004.

- [67] The RDFLib Team. rdfliib. <https://rdfliib.readthedocs.io/en/stable/>, accessed on: 14th December 2021.
- [68] Torsten Hothorn, Kurt Hornik, Mark A. van de Wiel, and Achim Zeileis. A Lego system for conditional inference. *The American Statistician*, 60(3):257–263, 2006.