INTERNATIONAL HELLENIC UNIVERSITY

SCHOOL OF SCIENCE AND TECHNOLOGY

MSC IN E-BUSINESS AND DIGITAL MARKETING

TITLE

Knowledge Graphs Tools and Applications

SUPERVISOR

DR. CHRISTOS BERBERIDIS

STUDENT

DIMITRIOS - PARASKEVAS TAGKOULIS

REG. NO.

3305180030

THESSALONIKI, 29/4/2021

Abstract

In this study, we analyze the definition, applications, and tools of Knowledge Graphs and highlight a use case for this technology in the medical sector. The study consists of two parts. First, the theoretical and second, the practical part. In the first part a literature review is conducted, and the main goal was to understand how Knowledge Graphs can be used by scientists in many fields. In the second part, we discussed about the methodology and evaluate some Knowledge Graph creation tools from the enterprise sector. Then, we used BioGrakn Covid to query and analyze large amounts of data and papers related to Covid-19. We run four queries of increasing complexity. First, *"Get all Genes associated with the Virus named "SARS"*, second *"Get all genes that encode proteins and their respective encoded proteins",* third "*Get all proteins associated with the virus named "SARS""*, fourth *"Get all genes that encode proteins and their respective encoded proteins that are associated with the virus named "SARS""* and fifth *"Get all genes that encode proteins and their respective encoded proteins that are associated with any coronavirus".* According to the theoretical and practical part, we can conclude that Knowledge Graphs can benefit enterprises by both saving research time, and by better understanding the information provided and the relations within through interactive visualizations.

Table of Contents

Abstract 2
Table of Contents 3
List of Figures 5
List of Tables 5
INTRODUCTION 6
CHAPTER 1: LITERATURE REVIEW 9
1.1 Introduction and Background 9
1.2 Critical literature review 15
1.2.1 Applications 15
1.2.2 Tools 20
1.3 Critical discussion 23
CHAPTER 2- METHODOLOGY 26
2.1 Literature review. 26
2.2 Deciding the suitable KPI's for this research. 26
2.2 Exploring the learning resources for each tool. 27
2.3 Evaluation of the learning curve. 28
2.4 Data used for Evaluation. 28
2.5 Performance evaluation. 29
2.5 Use Case. 31
CHAPTER 3- ENTERPRISE TOOLS EVALUATION 32
3.1 Tool Summary 32
3.1.1 TopBraid and Topbraid Composer 32
3.1.2 Grakn. 32
3.1.3 IBM Graph. 33
3.1.4 Amazon Neptune 33
3.1.5 Azure CosmosDB 33
3.2 Evaluation results 34
CHAPTER 4- KNOWLEDGE GRAPH USE CASE 35
4.1 The use case 35
4.2 Data sources 35
4.3 The Ontology 36

4.4 Results 37
CHAPTER 5- CONCLUSIONS AND REMARKS 42
References 43

List of Figures

Figure 1 Knowledge Graph Example, (Dwivedi, 2020). 10
Figure 2 Knowledge Graph Alignment Example, (Trisedia, Qi, & Zhang, 2019)..... 13
Figure 3 BioGrakn Covid Schema 36
Figure 4 Query 1: “Get all Genes associated with the Virus named “SARS”” 37
Figure 5 Query 2: “Get all genes that encode proteins and their respective encoded proteins” 38
Figure 6 Query 3: “Get all proteins associated with the virus named “SARS”” 39
Figure 7 Query 4: “Get all genes that encode proteins and their respective encoded proteins that are associated with the virus named “SARS”” 40
Figure 8 Query 5: “Get all genes that encode proteins and their respective encoded proteins that are associated with any coronavirus” 41

List of Tables

Table 1 Available Learning Resources Results..... 27
Table 2 Tool Learning Curve Results. 28
Table 3 Tool Loading Results. 29
Table 4 Tool 1-hop Query Results. 30
Table 5 Tool Scalability Test 31
Table 6 Final Evaluation Scores..... 34

INTRODUCTION

The modern era is characterized by the rapid digitization of all information, and the gradual change of the socio-economic environment, from analog to digital. In this context, many people increasingly rely on searching for information online, to cover their knowledge, to work, to carry out daily tasks, etc. In this environment, therefore, the source of information is especially important, and namely its accuracy and author. The value of correct and accurate information is of such an importance, that multiple organizations like the UN, EU, OSCE, and other have taken steps to protect its dissemination (Koulas, 2019), and some have even taken steps to address issues like misinformation (Koulas et.al, 2020). Due to the increasing phenomena of information recycling, databases are called to serve their "customers", giving them the best possible results, based on their search, however, in many cases, the results are recycled due to the multiple sources of information, and are therefore difficult to export (Pujara & Getoor, 2008).

Especially in some cases, the issue is complicated as the researcher (i.e. the one who seeks the information) is not able to recognize the accuracy and reliability of the sources from which he receives it. As a result, the researcher should spend more time at the intersection of his information, to avoid the re-transmission of incorrect information or the wrong impression of the information to him (Pujara & Getoor, 2008).

Additionally, in the effort of different companies to significantly reduce costs and risk from their operation, as well as in the effort of public and private entities to enhance their level of security and functionality, different tools are created to support these individuals. In particular, information technologies, telecommunications and various mathematical tools can be used to solve many of the problems found in the daily and special operation of businesses (Dwivedi, 2020; Nickel, Murphy, Tresp, & Gabrilovich, 2016).

Knowledge Graphs are tools that can help both to achieve the goal of optimizing the complex data management process and to reduce the risks associated with the information management process. As tools, they use structured data or even free text to make connections between information. Specifically, Knowledge Graphs are used by large companies such as Facebook ™, Google ™ and Wikipedia ™. Their use is observed in cases where it is necessary to categorize and correlate a

large amount of information, for example, the connection of a person with many events, or the connection of symptoms of a disease with a specific disease and so on (Berven, Christensen, Moldekev, & Opdahl, 2019; Heck, Hakkani, & Tur, 2013; Huang, Yang, van Harmelen, & Hu, 2017; Pujara & Singh, 2018; Rotmensch, Halpern, Tlimat, Horng, & Sontag, 2017).
This research attempts, on the one hand, to analyze the definition, applications, and tools of Knowledge Graphs and, on the other hand, to conduct an experiment which illustrates a use case for this technology in medical research.

In detail, the work is structured as follows:

This chapter is the introduction to the study. Its aim is a brief definition of the main topic and object of the work and the description of the methodology of both the literature review and the experiment on which the practical part of the work is based. The structure of the work is also included.

The first chapter of the work is the bibliographic review. The critical review is based on the study of a significant number of articles that have been published in the last decade and that have been published in reliable sources. These secondary sources are analyzed based on their content, while the analysis is divided into three parts: a) an introduction to the topic, b) a critical bibliographic review of specific articles in relation to the applications and tools of Knowledge Graphs and, c) a critical analysis of the points that need further investigation based on the judgment of the researcher.

The second chapter of this work is the description of the methodology used in this work. The methodology used is based on the corpora collected during the literature review phase of this study. In short the steps used for this study are the following: a) literature review, b) deciding the suitable KPI's for this research, c) exploration of the learning resources for each tool, d) the evaluation of the learning curve, e) finding data to be used for evaluation, f) performance evaluation g) selection of the tool and use case for this study.

The third chapter contains the evaluation of six enterprise tools that are used for the creation of Knowledge Graphs based on five key metrics, a) Load Speed, b) Query speed, c) Learning Curve, d) Learning resources, e) Scalability. The tools that will be evaluated are Topbraid, offered by TopQuadrant, IBM knowledge graph, Grakn offered by Grakn.ai, Neptune offered by Amazon and Azure Cosmos DB offered by Microsoft.

The fourth chapter includes the experiment in which, one of the tools that were evaluated on the third chapter, namely Grakn, and more specifically BioGrakn Covid, an open source knowledge graph to enable research in COVID-19 and related disease areas will be queried, evaluated and used to quickly spot relations that were identified by the academic research to highlight how such a tool can be used to drive research and business innovation forward.

The fifth chapter consists of the main conclusions and critiques of the paper as a whole. In detail, the purpose of this chapter is to describe the overall findings of the paper, its limitations and the author's suggestions for further research.

CHAPTER 1: LITERATURE REVIEW

This chapter critically analyzes a set of articles and studies on the importance of Knowledge Graphs as a tool in a set of areas. In particular, secondary sources that have been published after 2010 with the exception of one (1) article are analyzed. The aim is to make an analysis that is both relevant and informative to the reader. Also, a specific goal is to understand how Knowledge Graphs can be used by scientists in the field of finance and business, as well as in the field of information technology and security.

1.1. Introduction and Background

The method of Knowledge Graphs is a key part of the information sector, and specifically the field of artificial intelligence. It is a method of displaying and analyzing information, based on the correlations that develop between the individual elements of a data model. Utilizing the potential of this method is a key part of the modern science of data analysis, as it enables the further development of a data model and the extraction of additional information from it (Nickel, Murphy, Tresp, & Gabrilovich, 2016).

The specific field in which the method of Knowledge Graphs is used is the field of machine learning, which plays an important role in the development of artificial intelligence applications. In the case of these applications, certain criteria must be met so that the effective use of artificial intelligence can be demonstrated. A key factor in this context is the concept of logic, ie the ability of the machine to understand the information given to it and to create correlations. The above example with Spock and OWK is typical as in this type of information the result can be either vague or specific (Nickel, Murphy, Tresp, & Gabrilovich, 2016).

Databases operate in a way that allows the user to instantly locate the information they are interested in, based on data matching algorithms. In this way, when the researcher searches for a phrase in the database, it will be returned to a series of information, distributed according to the researcher's parameters. If the researcher adds additional parameters to his search, such as time period, origin, data type or information, etc., then the database will limit the displayed results. But

again the presentation of these will have a serial form. This method is very time-consuming in case the information sought must be characterized by absolute accuracy (Pujara & Getoor, 2008).

In this context, since 2012, Google, which is the largest information company in the world, has introduced Knowledge Graphs technology. This technology allows the search engine to return a series of basic data to the user, without him having to refer to their sources. This technology is particularly useful in cases where the information sought by the search engine operator is not very critical, but has been repeated many times on various websites, and is generally prevalent in search engine demographic data, or data shops and businesses (Paulheim, 2016).

However, this practice raises some questions, as this technology provides data directly, without guaranteeing their accuracy. The first and foremost question is: "what criteria are used to select the displayed data", which according to research is a combination of selecting common elements between the information, and selected information sources that are promoted by search engines. The second question that arises concerns the accuracy of this information, as in many cases the source is not mentioned. The answer to this question is the main object of this dissertation, where through the critical literature review and the application of the experimental procedure, the factors that compose the accuracy of the information are identified (Paulheim, 2016).

In general, Knowledge Graphs are based on identifying common elements as described in the following example (Dwivedi, 2020):

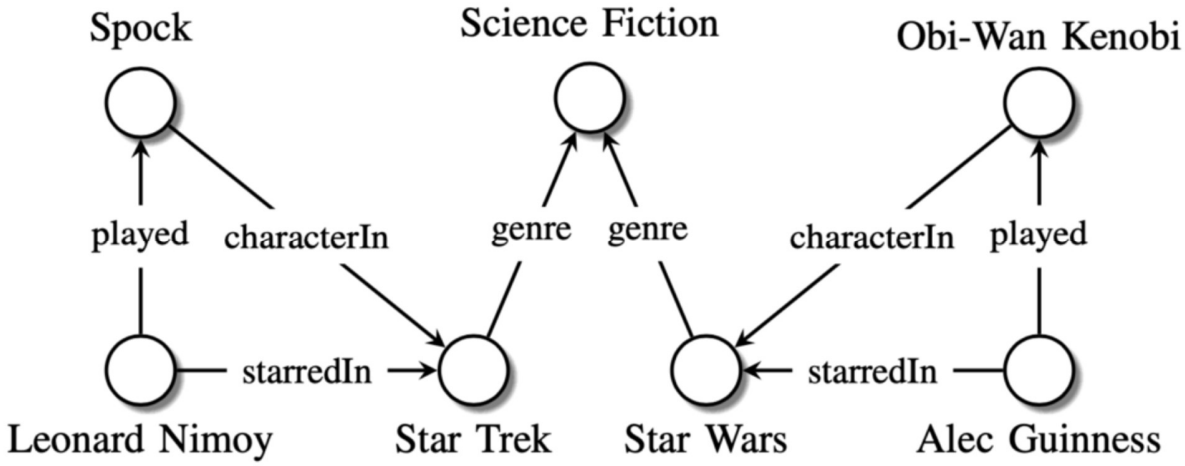

Figure 1 Knowledge Graph Example, (Dwivedi, 2020).

The example concerns two of the most famous characters in science fiction movies, namely Spock (Leonard Nimoy) and Obi-Wan Kenobi (Alec Guinness), who appeared in Star Trek and Star Wars, respectively. The characters in both the movies and later literature are diametrically opposed[1], and the two franchises have no connection, including the production studios. Based on the above, the correlation of these two characters would be impossible, and would not offer cross-referenced results, however, through the above Knowledge Graph, a correlation emerges as to the "genre" of the movies in which the two characters appear.

Based on the above, it appears that Knowledge Graphs go beyond the limits of a simple tool in the field of databases and can be used for big - data analysis. By exploiting these possibilities, the possibility arises to identify correlations between seemingly unrelated concepts. These possibilities are particularly important in problems that present themselves in the form of large volumes of data. Utilizing the capabilities of Knowledge Graphs, the researcher can identify elements that are related in ways that are not predicted without them, and then use the results to achieve his goal (Dwivedi, 2020).

At the same time, this technique can be the basis for defining problems, which were previously unknown as the correlations between the data, can present a new dimension to an existing problem, or indicate the factor that prevents the problem from being solved. In this context, it is particularly important to note that Knowledge Graphs are not related to the vague correlation of unrelated data, which are extensively used as arguments in conspiracy theories or false news. Instead, they are tools through which more dimensions of the elements of a set are identified, to indicate (if any) the correlations between them (Paulheim, 2016).

The basic technology behind Knowledge Graphs can be summed up to three main tools, which are (Pan, Vetere, Gomez-Perez, & Wu, 2017):

- Representation and reasoning of the information contained in a knowledge graph, and to an extent the databases.

[1] Both for those familiar with the characters, and those who are not with them, the two characters are based on different personalities, in one case Spock (Nimoy) is governed by logic and scientific facts, whilst the OWK (Guinness) acts based on the emotion and the collected wisdom of the years.

- Data storage, which comes in the form of the databases.
- Information Engineering, which can take various forms from methodologies and editors to design patterns.
- Knowledge learning, which is the main focus of the knowledge graph.

Based on the above, two "assumptions" are formed, the first is the closed world assumption (CWA), based on which the absence of correlations between elements of the model indicates an error (ie they do not exist), while the second is the open-world assumption (OWA), based on which the error is not certain in the absence of data. The difference between the two is particularly important as, in general, information is collected from open databases. As a result, the design of an algorithm or model, which will collect information from these databases, is feasible and is already widely implemented. However, in the case of CWA, the user assumes that the results are final, and therefore the correlations created are correct, while in OWA the user is not able to know if these results are final and if an error occurred during the construction of the knowledge graph (Nickel, Murphy, Tresp, & Gabrilovich, 2016).

Of course, as in most data analysis methods, in the case of Knowledge Graphs, there is a wide range of methodology and approach, to achieve the corresponding goal. In the research of Ulincy, (2016), The approach to the issue of Knowledge Graphs is the free capture of the correlations between the information, from the set of data provided. In practice, this method enhances the system's ability to form correlations between the elements of the set, without prior delimitation of permissible or non-permissible correlations. This method, as mentioned by the author, enhances the ability of companies to draw conclusions, based on general information that is free, without the need to specify the data received, especially personal data (Ulincy, 2016).

However, this approach has some drawbacks as without the delimitation of the permissible correlations, or the targeted correlations, the results of the research may be very vague or inaccurate, and in some cases the conclusions that will emerge may not be the answer to initial question that led to model construction and analysis (Ulincy, 2016). Another approach to the problem proposed by Trisedia, Qi, & Zhang, (2019), is the embeding of certain basic data analysis models, within the algorithms that are responsible for the construction of Knowledge Graphs, these

models, force the algorithm to look for correlations between the elements of the graph, even after its completion, significantly increasing its efficiency.

These models are described in this research as learning embendigs as they enhance the "learning" capabilities of the algorithm regarding the detected data. Specifically, the authors propose the following Alingment Model, which is used to increase the accuracy of the algorithm:

G_1
⟨lgd:240111203, geo:long, 11.3700843⟩
⟨lgd:240111203, lgd:population, 1595⟩
⟨lgd:240111203, rdfs:label, "Kromsdorf"⟩
⟨lgd:240111203, geo:lat, 50.9988888889⟩
⟨lgd:240111203, lgd:alderman, "B. Grobe"⟩
⟨lgd:240111203, lgd:country, lgd:51477⟩
...
G_2
⟨dbp:Kromsdorf, geo:long, 11.3701⟩
⟨dbp:Kromsdorf, rdfs:label, "Kromsdorf"⟩
⟨dbp:Kromsdorf, geo:lat, 50.9989⟩
⟨dbp:Kromsdorf, dbp:populationTotal, 1595⟩
⟨dbp:Kromsdorf, dbp:country, dbp:Germany⟩
⟨dbp:Kromsdorf, dbp:district, dbp:Weimarer⟩
...
Merged G_{1_2}
⟨lgd:240111203, geo:long, 11.3700843⟩
⟨lgd:240111203, :population, 1595⟩
⟨lgd:240111203, rdfs:label, "Kromsdorf"⟩
⟨lgd:240111203, geo:lat, 50.9988888889⟩
⟨lgd:240111203, lgd:alderman, "B. Grobe"⟩
⟨lgd:240111203, :country, lgd:51477⟩
⟨lgd:240111203, dbp:district, dbp:Weimarer⟩
...

Figure 2 Knowledge Graph Alignment Example, (Trisedia, Qi, & Zhang, 2019).

Through the above model the algorithm acquires the ability to apply the necessary connections, which lead to the construction of the graph. The application of the above can be done either in cases of local data processing (offline) from a finite set of data, or in cases of public data processing (online) where the data of the processed set are combined and processed, based on data located in public databases. Using the example of the introduction of the literature review, in the first case the researcher should have available a set of data which concerns multiple types of films, actors' names, film titles, type of role, etc. so that the algorithm proceeds to the necessary correspondences, in the second case, however, from the researcher's point of view, only the parameters of the analysis are required, and consequently the construction of the diagram (Kliegr & Zamazal, 2016).

The subsets also added to the knowledge graph construction algorithms also aim to solve other weaknesses of the original algorithms, which can lead to inaccurate display of information, or the loss of critical correlations that can change the final meaning of the result. In the research of Hamilton, Bajaj, Zitnik, Jurafsky, & Leskovec, (2018) this issue is a central object, in the sense of adapting the subsets, as complementary control conditions of the resulting results. Essentially, these are repetitive processes that examine correlations in order to identify "gaps" or "weaknesses" between them, which take the form of either the absence of a correlation or the existence of someone who can be described as vague or false (Hamilton, Bajaj, Zitnik, Jurafsky, & Leskovec, 2018). Similar results emerge from the research of Wang, et al., (2019) who examined the benefits of adding embeddings to a knowledge graph construction algorithm. According to the researchers, the main benefit was the increased ability of the algorithm to understand and "learn" the information, which significantly reduced the time frame for completing the process (Wang, et al., 2019).

However, the construction of a knowledge graph can prove difficult, through the method of collecting data from the Internet, in general, the companies that apply it first proceed to the "collection" of the necessary data, which are then categorized into clusters and then use the algorithm. for the construction of Knowledge Graphs. This process aims to filter out inaccurate or false information that is sometimes found on the Internet and is perhaps the most important part of building the right Knowledge Graphs that will show real correlations (Berven, Christensen, Moldekev, & Opdahl, 2019).

1.2 Critical literature review

Having analyzed above the definition of Knowledge Graphs, their use and their applications, at this point, one can deal with a critical literature review, which concerns the evaluation of the methods of application of Knowledge Graphs in the various scientific approaches in which they are used. For this reason, a number of contemporary articles are studied, which have been selected on the basis of the following criteria: a) their relevance to the specific subject of this research, b) to have been published after 2000, c) to have been published in reputable and recognized databases or journals for their validity and reliability.

In detail, in this literature review, the scope is to define:

a) The applications and use of Knowledge Graphs, meaning the various software and programs as well as sectors that make use of Knowledge Graphs.

b) The tools that are applied and their particular characteristics and uses.

1.2.1 Applications

Starting from the study of Popping (2003), in recent years there has been a particularly increased interest in the application and utilization of Knowledge Graphs in a number of areas that do not necessarily show any relationship between them. This is a phenomenon that is not common knowledge, but nevertheless it is a natural evolution due to the increase of IT applications in all areas of economic and daily life. Especially in the case of data analysis methods, many sectors are trying, in addition to taking advantage of existing methods, to innovate and to increase their influence in the respective sectors, and to receive the corresponding benefits (Popping, 2003).

Indeed, according to Popping (2003), the main field of application of Knowledge Graphs is data mining, i.e. the process of locating and extracting data from a seemingly infinite set of them. This

field is very promising over the last decade with its best-known technologies including blockchain, cryptocurrencies and various methods of data collection and utilization. However, the above field is also the main source of public concern regarding the collection and processing of personal data, which is mainly due to the lack of knowledge in the field (Popping, 2003).

Next, the paper of Kazeemi & Poole (2018) explains that Knowledge Graphs are a tool to illustrate the data and can cover the majority of areas. In particular, Knowledge Graphs can be created using both structured data and plain text. The data can, also, be either symmetrical or a-symmetrical, meaning that the use of Knowledge Graphs can alter a large number of restrictions that, typically, apply when studying and using facts and data.

In this context, then, Trivedi et al. (2017) explain that companies and services of all kinds try to take the lead in finding the best methods of data collection and processing, in order to fulfill their goals. Catering companies, for example, are trying to track down consumer trends while significantly reducing the cost of the tracking process. Companies and information services use the above technologies to analyze their information, and especially to verify it, so as to avoid sharing false information. It is also the main reason why today is often referred to as the "Digital Age" as more and more processes taking place on the planet use some form of digital technology, and consequently data processing (Trivedi, Maheshwari, Dubey, & Lehmann, 2017). In addition to this field, Rospocher, et al., (2016) report how this field can greatly benefit from the use of Knowledge Graphs, as they can combine real-time information generated by information organizations, and available historical data to cover in depth a fact (Rospocher, et al., 2016).

On the other hand, as Heck et al. (2013) note, this does not make the method of collecting information, or the use of Knowledge Graphs immoral, as this is a practice which (at least in the EU) is highly controlled by formal mechanisms, and by the legal framework. It is also a natural change in the field of advertising, due to the increasing conversion of users from traditional media (and therefore the main channels of advertising material) to the Internet, which allows companies to have a more “direct” user contact. Also, the user is significantly more likely to respond positively to ads that interest him, thus significantly increasing the efficiency of the advertising sector. The above is just one example of the use and utilization of Knowledge Graphs, as the capabilities of this technology can (theoretically) be applied in any field that uses information systems (Heck, Hakkani, & Tur, 2013).

Besides, a major confusion observed in the issue of Knowledge Graphs is the issue of their capabilities. Knowledge Graphs consist of the correlations generated between the elements of the data set provided in the algorithm. They are essentially a "map" that shows the relationships and commonalities between seemingly different elements. What they cannot do is produce 100% accurate conclusions or predictions depending on the change of data, as they do not have this ability. The reason is the fact that the data that the algorithm is called to process, and the correlations are shown by the knowledge graph are real, that is, they have been generated, located, constructed, or recorded. For this reason, this technology is categorized as a "tool" and not a separate field of data processing. However, this limitation can be bypassed if this technology is combined with other data analysis methods, which include predictive analytics (Xie, Liu, & Sun, 2016).

In this case, the "history" of the data, which is revealed by the Knowledge Graphs, is processed by the tools for predicting the evolution of the data, in order to produce results related to their change. The stock market is the most representative case of this. Since the course of the price of a share depends on a wide range of factors related to the internal and external environment of the company, it is difficult to accurately or approximately predict the price change, solely using predictive analytics tools. In this context, the use of Knowledge Graphs can prove to be particularly critical, as it can combine a particularly large set of data, on the past price changes, but also the external factors that prevailed during this period, and present them, and then through the use of predictive analytics to assess the course of the stock in the immediate future (depending on the method or the tools) (Guo, Wang, Wang, & Guo, 2017).

Based on the above, and according to Dwivedi (2020) the possibility of using Knowledge Graphs is very vague, which can cover any possible field that uses even basic data analysis tools. Some examples of this are the field of public safety and law enforcement, the field of production and sale of products, the financial sector, the information sector, the mathematics and physics sector, the medical and pharmaceutical sector, and many others. Of course, the practices vary depending on the form of the market, the technological level of each country, the value of each sector in the respective market, etc., however in recent years this technology has become more and more a key tool in information analysis, especially in countries in Europe and North America, but also in other regions such as Russia, Latin America and Asia (Dwivedi, 2020).

Especially in the IT fields, the use of Knowledge Graphs is a regular practice in big data analysis as in these cases the traditional tools often do not meet the needs of the problem, resulting in their results either excluding a percentage of the aggregate data, or present inaccurate results due to time constraints on resource consumption constraints. The use of Knowledge Graphs, unlike traditional methods, presents the patterns that form the elements of the set, based on certain parameters by the analyst, in order to "show" the set that best meets the needs of the problem. Essentially, this method makes it possible to isolate a subset of data, whose data show the correlations that the researcher is looking for, and then the effort to complete the analysis focuses on these points (Kliegr & Zamazal, 2016).

Additionally, further analyzing the applications of Knowledge Graphs, according to Nickel et al. (2016), it is found that, in no case does this method contain a degree of error, as the volume of data increases, the resources required to create the knowledge graph increase significantly. This consumption can sometimes lead to interruption of the process, as the costs outweigh the benefits. For this reason, Knowledge Graphs construction algorithms usually include "safety valves" which refer to either the number of iterations of the process or the completion time frame. Due to the ability of algorithms to look for increasingly vague correlations between elements, these valves are necessary in order to avoid the possibility of an infinite loop. As a result, before starting the process, the user defines the time frame based on the available computing resources, while in some cases it can limit the "depth level" to which the algorithm will proceed to create a correlation (Nickel, Murphy, Tresp, & Gabrilovich, 2016).

This level based on the example of the introduction is the creation of new nodes that respond to each correlation. For example, if more than two nodes are required to create the correlation, then the correlation can be considered vague and therefore not included in the knowledge graph, this, of course, depends on the researcher himself and the tools he uses to complete the procedure, as well as the type of problem (Popping, 2003; Arenas, Grau, Kharlamov, Marciuška, & Zheleznyakov, 2016).

Coming to specific applications of Knowledge Graphs, one can define a number of uses that are more elaborate and case – specific. For instance, as Rotmesnch et al. (2017) argue, that, to a very large extent, Knowledge Graphs can be used in health institutions and organizations. In detail, in

this specific instance, Knowledge Graphs are preferable to manual processing and analysis of large databases and are, additionally, more precise, organized and clear to the user.

As one can assume, then, due to the fact that Knowledge Graphs offer a significant advantage in cases where associations of different factors are necessary, the creations of such graphs can limit the amount of time required to process a case. In the health sector, speed, accuracy and the possibility to get the full image in a certain instance can be life - saving. Similar applications can be found in health when it comes to the function of specific machinery and medical instruments as well as managing information online in relation to specific diseases and symptoms etc. Therefore, studying and creating Knowledge Graphs can have an impact on public health and public safety (Rotmensch, Halpern, Tlimat, Horng, & Sontag, 2017).

This issue is also the subject of research by Huang, et. al., (2017) who examine the capabilities of Knowledge Graphs technology in order to improve treatment methods for diseases such as depression, stress, etc.. This is a particularly important piece of technology as it opens a promising sector, for the expansion of its applications. Specifically, the researchers are considering the possibility of making Knowledge Graphs to treat specific diseases (in the case of research, the goal is to treat depression, but this is a case study). The researchers report that by using Knowledge Graphs, the parameters of the procedure can be the patient's symptoms, and the databases the existing medical research libraries, so that a graph can be obtained that will not only respond to the disease, but also to the particularities of the patient, significantly increasing the efficiency of the therapeutic process (Huang, Yang, van Harmelen, & Hu, 2017).

Moreover, in the paper of Krompaß, Baier, & Tresp (2015) one can understand why Knowledge Graphs are essential in production. In detail, the researchers explain that, due to the fact that machines use data that are organized and coded, for instance structured databases, it is crucial that the information necessary is categorized. Additionally, Knowledge Graphs make use of statistics and the relationships created can be used to predict and assess phenomena, relations and impacts of different factors on others. In the following subsection, “tools” the different ways that these structures can be formed are explained.

1.2.2. Tools

In general, Knowledge Graphs are key tools, which enhance the ability of analysts to identify correlations that would not otherwise be visible. Depending on the scope, however, the process for constructing a knowledge graph takes a different form, as the specifics of each sector, and consequently the data that the algorithm is required to process, must be taken into account to complete the process. In this context, many analysts use a number of basic algorithms or bases in order to prepare the basic "trunk" of the process. These bases, on a case-by-case basis, concern the needs of the analysts and the goal they have set for the construction of the knowledge graph. The base is then expanded according to the specifics of the sector or the problem to be solved.

Therefore, having considered the above, the available tools and mathematical tools are analyzed which are used for the design and implementation of Knowledge Graphs based on the available literature. So, starting with the article by He, Liu, Ji, & Zhao (2015), Knowledge Graphs can be represented in the Gauss space and by using complex mathematical models. The main tool in this case is modeling using statistical models to measure certainty and uncertainty. For this purpose, the user / researcher uses a Gaussian distribution and aims to find the mean and then calculate the variance and covariance. The logic of these models is that the relationships / correlations between the variables are calculated. Depending on how far away from the average (variation) the various "points" (facts) are, the more one can calculate the certainty and uncertainty, the correlations and the evolution, by grace, of a phenomenon. This process, however, is very complex, and requires the use of specialized mathematical tools and great computing power (He, Liu, Ji, & Zhao, 2015).

In a rather more explanatory study, Voskarides et al. (2015), explain how a user can extract information online and use annotated text to create a knowledge graph. The researchers used manually entered annotated phrases and sentences to discover the main challenges associated with the use of text in the creation of a knowledge graph. Their methodology of analysis has as such:

- Organization of sentences according to their length as well as the various text features that are important in the analysis.
- Definition of the number and types of entities used.
- Definition of the relationship features.

- Explanation of the characteristics of the sources and the position of the sentences.

As a matter of fact, the selected methodology in the case of the paper of Voskarides et al. (2015) is rather specific. To define the density of the text, the authors used a “density” formula and to define the relationships and the links between factors they used linear equations. In this paper it has been proven that the main challenge is, perhaps, the ranking of sentences and of the entities rather than finding and explaining the relationships between the parameters and the entities. The conclusions of this paper, therefore, are, to a large extent, important for machine learning and software design in general and not only for the creation of Knowledge Graphs (Voskarides, Meij, Tsagkias, De Rijke, & Weerkamp, 2015).

In the research of Rospocher, et al., (2016) the issue of constructing Knowledge Graphs in the field of news, requires the use of tools such as ECKGs (Event - Centric Knowledge Graphs) which utilizes the methods of natural language and web semantics, in order to construct graphs that include information from many sources, and at high speed, two features which are crucial in the field of information. According to the researchers, this process can be a significant change in the field of media, as the news can now undergo dynamic changes, as events change, while additional data can be revealed, which under other circumstances they would be hidden (Rospocher, et al., 2016), as used by Befa, M. & Kontopoulos, E. & Bassiliades, N. & Berberidis, C. & Vahavas, I. to power a university CMS that could dynamically search and navigate through a Knowledge Graph, via semantic queries, to retrieve fragments and render them as interactive HTML (2010).

In part, this dynamic data collection is a significant advantage for organizations and services whose processes depend on the accuracy and speed of the results produced. However, these practices raise serious questions about the origin of this data and information, especially when it comes to protecting the personal data of internet users. This issue is the main object of research of Qian, Y., Zhang, & Chen, (2016), who examine the possibilities of using Knowledge Graphs, with the aim of detecting malicious comments, online bulling, etc. In their research they claim that the process they propose is aimed at identifying and determining the motives of the attacking parties, in order to assess the situation and take appropriate action. It is a deanonymization process that aims to expose people who, under the guise of anonymity offered by social media, attack people, often with malicious intent (Qian, Y., Zhang, & Chen, 2016).

The above process includes another method which is often omitted in reports related to the construction of Knowledge Graphs and is the process of text recognition or textual information. This process, which is also the subject of research of Wu, et, al., (2016), which concerns the ability of the algorithm to recognize patterns within written language, and to translate them into information or elements which are then used to construct the graphs (Wu, Xie, Liu, & Sun, 2016).

Then, regarding the way of using and, consequently, the tools for the creation and use of Knowledge Graphs, Pujara and Singh (2018) refer to the possibility of extracting data for the creation of Knowledge Graphs from static texts. Specifically, they describe in detail and step by step this process as follows (Pujara & Singh, 2018):

- Select a text that includes information that need to be sorted and / or codified
- Extract information from the annotated text using a form of "keywords", i.e. names and places (cities, locations, countries)
- Create connections between the people and the locations

In detail, Pujara & Singh (2018) note that:

a) verbs act as connections. This means that verbs will be used to create the knowledge graph and define the connections between, for instance, the person and the location

b) articles, pronouns and prepositions need to be also defined to act as "tokens" to create paths.

c) linking may be challenging when it comes to people with the same name and or locations that also share a same name (i.e. John). Then, due to the fact that the linking should be coherent, further connections can be traced (i.e. George + Washington but not George + Washington + DC).

Moreover, the researchers clarify that the Knowledge Graphs can be closely supervised, "semi" – supervised or not supervised. This means that, depending on factors such as capacity, skill, the urgency, the need for effort and the importance of the information and the overall process, each programmer and user can choose to engage closer or less close to the process (Pujara & Getoor, 2008).

Additionally, Lee et al. (2018) attempt to explain the use of the tools available to create and comprehend the use of a knowledge graph, using more complex terms and explanations. In detail, the researchers follow a similar model to Pujara and Singh (2018), in the sense that they start by explaining how one can import data and then, how the aforementioned connections are made. In detail, they explain that one can define some positive and some negative connections (i.e. sunny + weather positive), and classify the connections and methods using patterns that are similar in everyday interactions. Moreover, the creation of Knowledge Graphs appears to follow the methodology of creating an empirical model, in the sense that one starts from a central point and, then, using multiple hypotheses, connects different variables with one another (Lee, Fang, Yeh, & Frank Wang, 2018).

Moreover, Choudhury, et al. (2017) discuss the process of creating Knowledge Graphs using the NOUS method. NOUS is derived from the Greek word “knowledge” or “thought” and is a framework that can be briefly explained as such: a) using a comprehensive and systematic model to curate information in a knowledge graph, b) a tool to discover trends and relations among data. In this framework, the process starts by collecting scalable data from different sources such as papers, text documents and websites. Then, the user can start processing the documents / facts / data by discovering relationships among them and to dis-ambiguate them. Following, they can create patterns and define certain “rules”.

1.3. Critical discussion

One of the most important conclusions of the literature review and the theoretical part of the study is that it can be particularly difficult to define specific "applications" for the construction of Knowledge Graphs, especially because they cover a very wide range of services and areas where they can be used, but also because of the vagueness that often prevails in describing these applications. The most well-known and regular form of Knowledge Graphs, and at the same time the main form that all users of electronic devices encounter is advertising, and specifically targeted advertising. Through Knowledge Graphs companies have the ability to provide users with targeted

advertising content according to their preferences, the main condition is the consent of the user for this process when entering a website.

Additionally, it is important to underline that many researchers argue that existing methods of building Knowledge Graphs can be characterized by inaccuracies and limitations, as companies that use existing construction methods often limit the possibilities of the algorithm to search data in databases in order to save resources and reduce the cost of the process. In their research, for instance, Paulheim, (2016) this issue is the focus of the study, as the various methods of making graphs are evaluated, as well as their evaluation methods. Specifically, the methods that are examined and present the best results are (Paulheim, 2016):

- Freebase, which is an algorithm that has access to more than 50 million databases, and 8 billion data. It is one of the largest software building Knowledge Graphs and is often used in information and training applications.

- Wikidata, which is also one of the largest Knowledge Graphs builders and recently acquired Freebase.

- Google's knowledge graph is the first with this name as Google established and patented the specific name for the process. Google, being the largest database in the world, offers the possibility to this software, to access practically infinite numbers of data and data, for the construction of Knowledge Graphs.

The following example is given for a deeper understanding of this:

1. Suppose a user enters a political information website. His website requests permission to send small packets of information, which in turn lead to the development of correlations, and are commonly known as cookies.

2. The user clicks accept and then the website has access to information that a) is not protected by any legal framework of the country to which the user's IP corresponds, and b) is targeted by the company as critical in determining the advertising material it will provide to the user. Regardless of the user's time on the site, the company responsible for the process has collected

the time spent, the links chosen (in the case of news sites the articles), the time spent on each of them, and their reactions which are all used in the promotional material.

Through the above information, a knowledge graph is formed which makes the necessary correlations. In this way, specific clusters are formed, each of which corresponds to a user category.

Additionally, if one attempts to give a more precise example based on the overall literature review, as Pujara & Singh (2018) noted, one can define that a person (John + Lennon) is associated with an entity (Beattles) and use verbs ("plays", "belongs", "acts", "likes", "left", "hates") to create the positive and negative associations that Lee et al. (2018) define. Next, the user can use different queries to find and categorize the relations and visualize data after export as Koukaras, Berberidis & Tjortjis noted (2020). Last, the information must be scalable, so different software can be used to build the final Knowledge Graphs and present the data extensively. Last, Choudhury et al. (2017) explain that, particularly when speed is more important, one can use fewer parameters to create structures, i.e., for security purposes. However, it can be summarized that, depending on the scope, the level of complexity, the size of data as well as the user's capabiltiy, different methods, tools and applications can be selected.

Particularly with regard to complexity, one can also note the work of Jayaram, Khan, Li, Yan, & Elmasri (2015) who explain that, as knowledge and data are limiteless, the creation of Knowledge Graphs may be considered a necessity, due to the fact that they allow the users to explain a number of questions by creating associations. This means that the users, one the one hand, can access an overview of the facts and data and, at the same time, answer complex questions by studying the data without having to ask clear questions. The main application, though, of Knowledge Graphs is that they can be used by non-experienced users and can be used in more than one area.

Therefore, if within the same website a user searches exclusively for news related to the car and omits the rest, then respectively the ads will target his preference if, for example, he is only interested in political articles, then the correlation between the content of the article is examined, and the length of stay within the link.[2]

[2] Note: Besides, this policy does not apply exclusively to specific websites, as many of the companies related to advertising cooperate with more than one, with the result that when it enters other websites, the advertising material

CHAPTER 2- METHODOLOGY

2.1 Literature review

The methodology followed for construction the corpora for this literature review was simple. Only papers published from 2010 and onwards were included as we wanted to have a recent and updated dataset and to examine the recent and emerging methodologies that have been used by the researchers the past years in order to construct, evaluate and improve Knowledge Graphs.

Google Scholar was used to apply the date filter and to search for papers matching a

combination of the major keywords such as “Knowledge Graphs”, “Graph databases” and

“creation” or “evaluation”.

From the query results papers were selected after reading the abstract sections and skimming through the rest of the paper from the first three (3) pages of the results with a priority given to the most cited ones.

2.2 Deciding the suitable KPI’s for this research

After conducting the literature review it was apparent that there was little material covering the performance of enterprise level solutions that can be used to create, manage, and use Knowledge Graphs. For this reason, the KPI’s that were picked for this study were Dataset Load Speed,

---

concerns the previous one. This is more understandable for a user when he enters a website selling electronic devices to evaluate which one he wants to acquire. Even if he does not complete the purchase, there is a possibility that the same devices will be advertised to the user on the next web page.

Query speed, Learning Curve for each tool, Learning resources available for each tool and Scalability for each tool.

2.2 Exploring the learning resources for each tool

For this part of the study, each possible learning resources piece published by the respective tool's creator was considered. Such material could be either official documentation, whietpapers, official tutorials, public video tutorials and either live or prerecorded webinars offered by the creator. After locating and evaluating each tool's corresponding material the following table with the corresponding score was created.

Learning resources						
Product/Material	Documentation	Whitepapers	Blog form tutorials	Video tutorials	Webinars	**Total**
Topquadrant / Topbraid	X		X			**2**
IBM Graph	X		X	X		**3**
Grakn	X	X	X	X	X	**5**
Amazon Neptune	X	X	X	X	X	**5**
Azure Cosmos DB	X		X	X		**3**

Table 1 Available Learning Resources Results

2.3 Evaluation of the learning curve

Using the available learning resources, each tool's learning curve was estimated by measuring the time needed from installation (in the cases of Topbraid and Grakn) to having loaded and queried the Graph500 data. It is worth noting that the three cloud-based tools (IBM graph, Amazon Neptune and Azure Cosmos DB) use similar workflows and infrastructure, resulting in a smoother learning curve.

Learning Curve *1 (Extremely Steep) - 5 (Smooth)*				
Topquadrant / Topbraid	IBM knowledge graph	Grakn	Amazon Neptune	Microsoft Azure Cosmos DB
1	4	5	4	4

Table 2 Tool Learning Curve Results.

2.4 Data used for Evaluation

The dataset used for the evaluation is generated from Graph500 (http://graph500.org) and taken from The Network Data Repository (Ryan A. Rossi and Nesreen K. Ahmed, 2015). The data used from the repository is the Graph500-scale18, which contains 7.6 million edges, as a tab-separated edge list like the following:

U1 U3

U1 U4

U2 U3

The data is then formatted as a csv for enhanced compatibility. There is no need for a separate vertex list since there are no attributes on the Graph 500 data.

2.5 Performance evaluation

Experiment setup:

- Performance evaluation was conducted on Windows 10, with intel i5-8250U 1.60GHz quad core 4 nm 8 thread processor with 8GB RAM.
- All cloud-based tools configured at their base options with Gremlin as the Graph API of choice.
- Data loaded as csv for all tools and conversion times or schema creation times (Grakn, Topbraid) were not included in the calculation since they are dependent on the user's familiarity with the platform.

Loading the dataset in each tool is repeated 10 times. The procedure is pretty straightforward for every tool, especially for the three cloud based ones since the procedure is similar. For Grakn specifically the data has to be loaded to the workbase through a client. For this study the Python client is used.

Loading Time *1 (Slowest) - 10 (Fastest)*					
	Topquadrant / Topbraid	IBM knowledge graph	Grakn	Amazon Neptune	Microsoft Azure Cosmos DB
Time (ms)	89500	4997	60092	10695	3920
Score	2	8	4	7	9

Table 3 Tool Loading Results.

To evaluate graph traversal performance, 1-hop-path neighbor count query is used. This query asks for the total count of the vertices which have a 1-length path from a starting vertex. For each tool we measure the query response time for the query “Count all 1-hop-path endpoint vertices for 300 fixed random seeds”.

Query Time *1 (Slowest) - 10 (Fastest)*					
	Topquadrant / Topbraid	IBM knowledge graph	Grakn	Amazon Neptune	Microsoft Azure Cosmos DB
Time (ms)	Failed	2.6	5.7	1.79	3.1
Score	1	8	5	9	7

Table 4 Tool 1-hop Query Results.

After the loading and querying tests the dataset is split into three parts and the parts loaded simultaneously to gauge the scalability of each tool. The slowdown percentage is calculated based on the previous loading time.

Loading Time as three instances *1 (Most slowdown) - 10 (Least Slowdown)*					
	Topquadrant / Topbraid	IBM knowledge graph	Grakn	Amazon Neptune	Microsoft Azure Cosmos DB
Time (ms)	Failed	5448	67150	12221	4190

Slowdown (%)	-	9.03	11.75	14.27	6.89
Score	1	7	6	4	9

Table 5 Tool Scalability Test

2.5 Knowledge Graph Use Case

For this part of the study, we will be using Grakn and more specifically BioGrakn Covid, which a knowledge graph built using Grakn by its creators, by integrating Covid-19 research with other publicly available data sources. It enables users to query all that knowledge using Graql, which is highly natural and intuitive. The queries selected for this use case are of increasing complexity to highlight the iterative nature of using such tools as knowledge bases to guide decision making, research and other business or academic tasks. The queries used are the following:

First, *"Get all Genes associated with the Virus named "SARS"*, second *"Get all genes that encode proteins and their respective encoded proteins",* third "Get all proteins associated with the virus named "SARS", fourth *"Get all genes that encode proteins and their respective encoded proteins that are associated with the virus named "SARS"* and fifth *"Get all genes that encode proteins and their respective encoded proteins that are associated with any coronavirus".*

CHAPTER 3- ENTERPRISE TOOLS EVALUATION

3.1 Tool Summary

In this section, five tools will be evaluated for their knowledge graph creation capabilities:

- TopBraid offered by TopQuadrant
- Grakn offered by Grakn.ai
- IBM Graph offered by IBM
- Amazon Neptune offered by Amazon
- Azure CosmosDB offered by Microsoft

3.1.1 TopBraid and Topbraid Composer

TopQuadrant's TopBraid Suite provides TopBraid Composer as a modeling environment that allows the user to make and manage domain models and ontologies within the Semantic Web standards RDF, RDFS and OWL. Composer is an ontology editor and knowledge-base framework that has visual editing support similarly as interoperability with UML, XML Schema and databases. TopBraid Composer is made on the Eclipse platform and Jena API. Testing, consistency checking and debugging is supported by built-in OWL Inference engine, SPARQL query engine and Rules engine.

3.1.2 Grakn

Grakn is constructed using several graph computing and distributed computing platforms,such as Apache TinkerPop and Apache Spark. Grakn is intended to be sharded and replicated over a network of distributed machines. Grakn uses a labelled, directed hypergraph as its underlying organisation. Grakn allows users to declare entities, resources, relations, and roles as an ontology. It uses its own graph command language Graql, which is declarative, knowledge-oriented and uses machine reasoning to retrieve explicitly stored and implicitly derived knowledge.

3.1.3 IBM Graph

IBM Graph is a fully managed property graph-as-a-service that allows the user to store, query and visualize data points, connections and properties. Built with the Apache Tinkerpop graph analytics framework it can transform and optimize gremlin queries into SQL statements, which get efficiently processed in IBM Db2 over a JDBC connection. It works by creating a virtual graph through a Graph overlay file that defines each row during a table as either a vertex or a position.

3.1.4 Amazon Neptune

Amazon Neptune is a fully managed graph database service. It allows users to simply build queries that efficiently navigate highly connected datasets, with popular graph models Property Graph and W3C's RDF, and their respective query languages Apache TinkerPop Gremlin and SPARQL.

3.1.5 Azure CosmosDB

Azure Cosmos DB built by Microsoft is a proprietary globally-distributed, multi-model database service "for managing data at planet-scale". it's schema-agnostic, horizontally scalable and customarily classified as a NoSQL database. To enable Graph storage and traversals it uses Azure Cosmos DB Gremlin API, Azure’s version of the favored graph traversal language of Apache TinkerPop.

3.2 Evaluation results

Finally, after running all the tests mentioned above, the final score for each tool is calculated by summing each respective tool's score per metric. The final results as well as the results for each test can be seen in Table 6.

Final Evaluation Scores					
Tool / Metric	Topquadrant / Topbraid	IBM knowledge graph	Grakn	Amazon Neptune	Microsoft Azure Cosmos DB
Load Speed	2	8	4	7	9
Query speed	1	8	5	9	7
Learning Curve	1	3	4	3	3
Learning resources	2	3	5	5	3
Scalability	1	7	6	4	9
Total	**7**	**29**	**24**	**28**	**31**

Table 6 Final Evaluation Scores

CHAPTER 4- KNOWLEDGE GRAPH CREATION USE CASE

4.1 The use case

For this study we will be using BioGrakn, Grakn's Semantic Database for Biomedical Sciences and more specifically BioGrakn Covid, the open-source knowledge graph created by Grakn.ai by using Covid-19 related research. This tool can query and analyze large amounts of data and research papers related to the Covid-19 virus, speeding up the research to cope with the virus and return to normality. It enables its users to quickly trace sources and identify articles and the information therein as well as visualize relations identified in the corpora.

4.2 Data sources

Currently the Covid Knowledge Graph is populated using data from the following sources as listed by the Grakn.ai team:

1. CORD-19: The original corpus which includes peer-reviewed publications from bioRxiv, medRxiv and others.
2. CORD-NER: The CORD-19 dataset that the White House released has been annotated and made publicly available.
3. Uniprot: The team downloaded the reviewed human subset, and ingested genes, transcripts and protein identifiers.
4. Coronaviruses: This is an annotated dataset of coronaviruses and their potential drug targets put together by Oxford PharmaGenesis based on literature review.
5. DGIdb: The team has taken the Interactions TSV which includes all drug-gene interactions.
6. Human Protein Atlas: The Normal Tissue Data includes the expression profiles for proteins in human tissues.
7. Reactome: This dataset connects pathways and their participating proteins.
8. DisGeNet: Curated gene-disease-associations dataset, which contains associations from Uniprot, CGI, ClinGen, Genomics England and CTD, PsyGeNET, and Orphanet.
9. SemMed: This is a subset of the SemMed version 4.0 database, about genes included in the CORD_NER dataset.

4.3 The Ontology

Ontology is Graql's formal specification of all the relevant concepts and their meaningful relations in the use case domain. It must be defined in order to load data to the Graph. The schema allows objects and relationships to be classified into distinct types, enabling automatic reasoning, such as inference (extraction of implicit information from explicit data) and validation (discovery of inconsistencies in the data). Grakn ontologies use four concept types for modeling domain knowledge. The classification of concept types is made by declaring every concept as a subtype of one of the four available concept types: entity, relation, role, and resource.

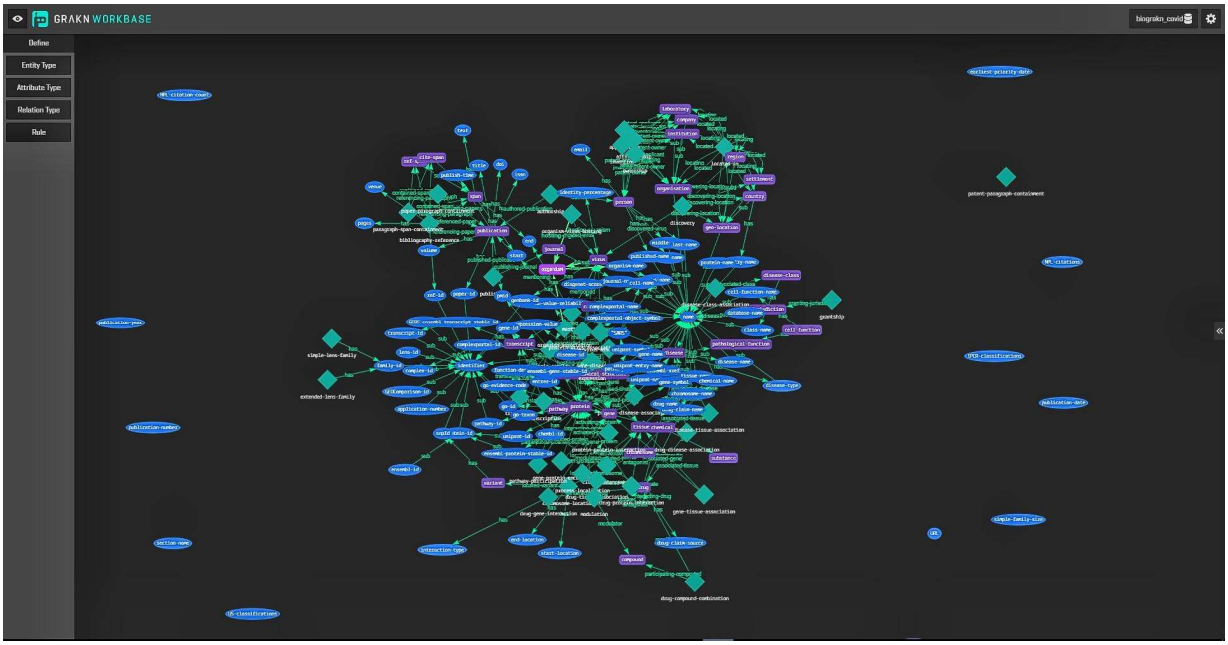

Figure 3 BioGrakn Covid Schema

4.4 Results

In this section, some representative queries of advancing complexity and results for typical Covid related problems are presented.

Query 1: “Get all Genes associated with the Virus named “SARS”

Since the main subject of interest in the corpora is the SARS – COV virus, we will be searching for this virus and all genes associated with it in the corpora. In order to find the associated genes, we query for the gene - association relation, which points out all the related entities, from which we extract the genes associated with SARS, printing their symbols and names. The following Graql query returns the desired results, shown in Fig. 4 in graph form:

match

$v isa virus, has virus-name "SARS";

$g isa gene;

$1 ($g, $v) isa gene-virus-association; get; offset 0; limit 100;

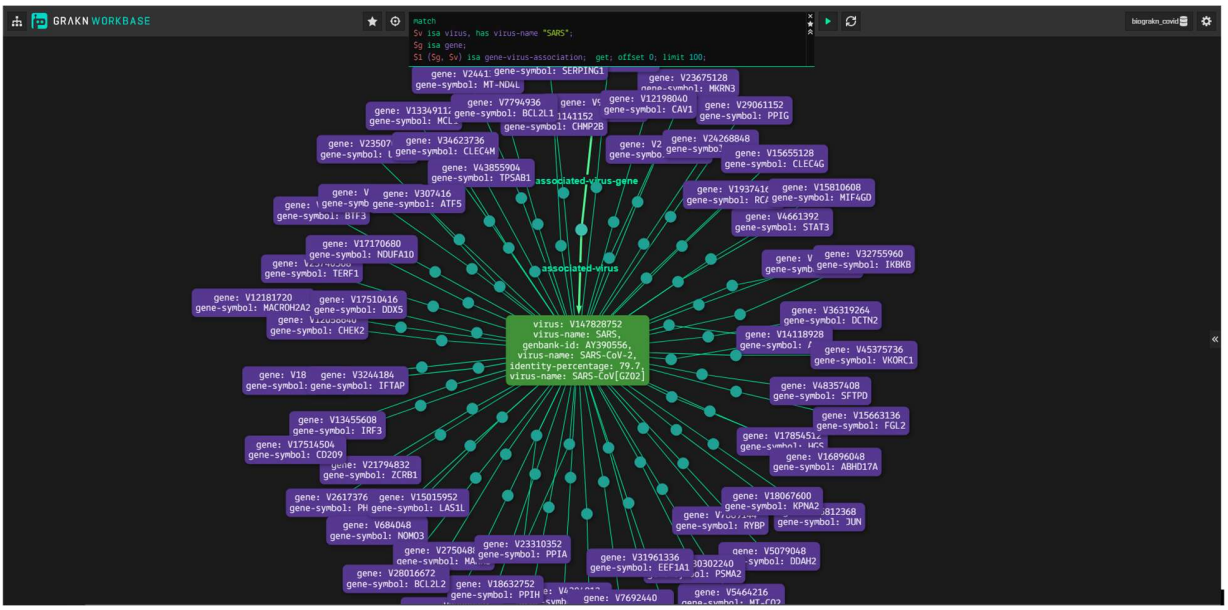

Figure 4 Query 1: “Get all Genes associated with the Virus named “SARS”

Query 2: “Get all genes that encode proteins and their respective encoded proteins”

The second query is used to highlight a second relation type, the gene-protein-encoding relation, which identifies the genes that encode proteins and their respective encoded proteins. The following Graql query returns the desired results, shown in Fig. 5 in graph form:

match

$g isa gene;

$p isa protein;

$1 (encoding-gene: $g, encoded-protein: $p) isa gene-protein-encoding;

get; offset 0; limit 100;

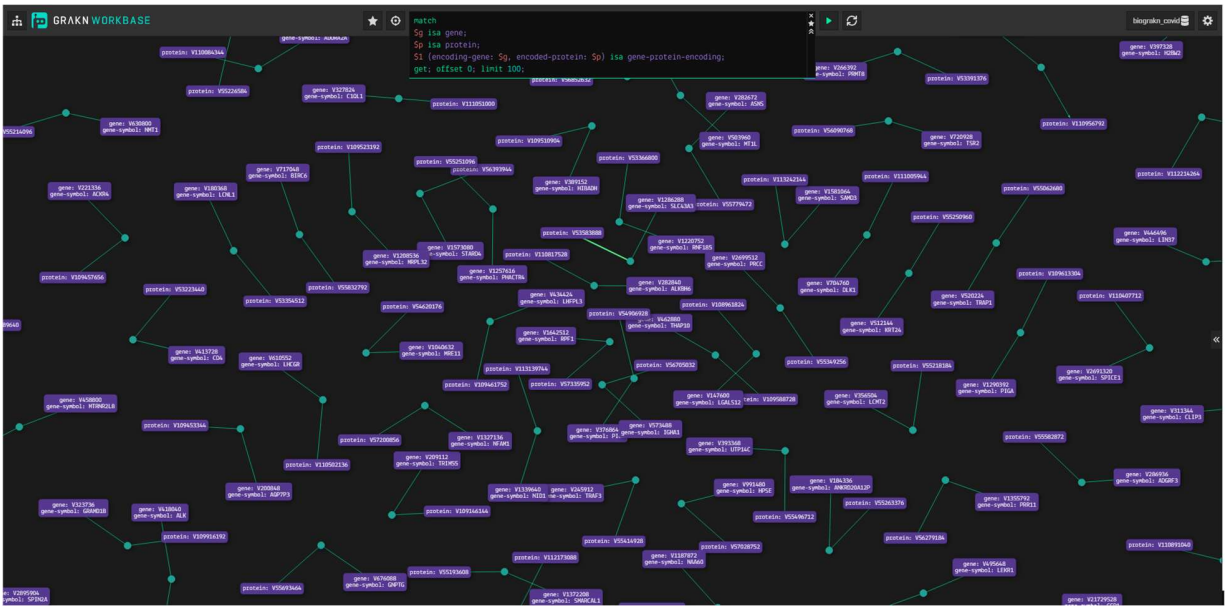

Figure 5 Query 2: “Get all genes that encode proteins and their respective encoded proteins.”

Query 3: “Get all proteins associated with the virus named “SARS”

The third query identifies a third relation type, the protein-virus-association relation, which identifies the proteins that are associated with viruses in the corpora. This association points out all

the related proteins, from which we extract the ones associated with SARS, printing their symbols and names. The following Graql query returns the desired results, shown in Fig. 6 in graph form:

match

$v isa virus, has virus-name "SARS";

$p isa protein;

$1 ($p, $v) isa protein-virus-association;

get; offset 0; limit 100;

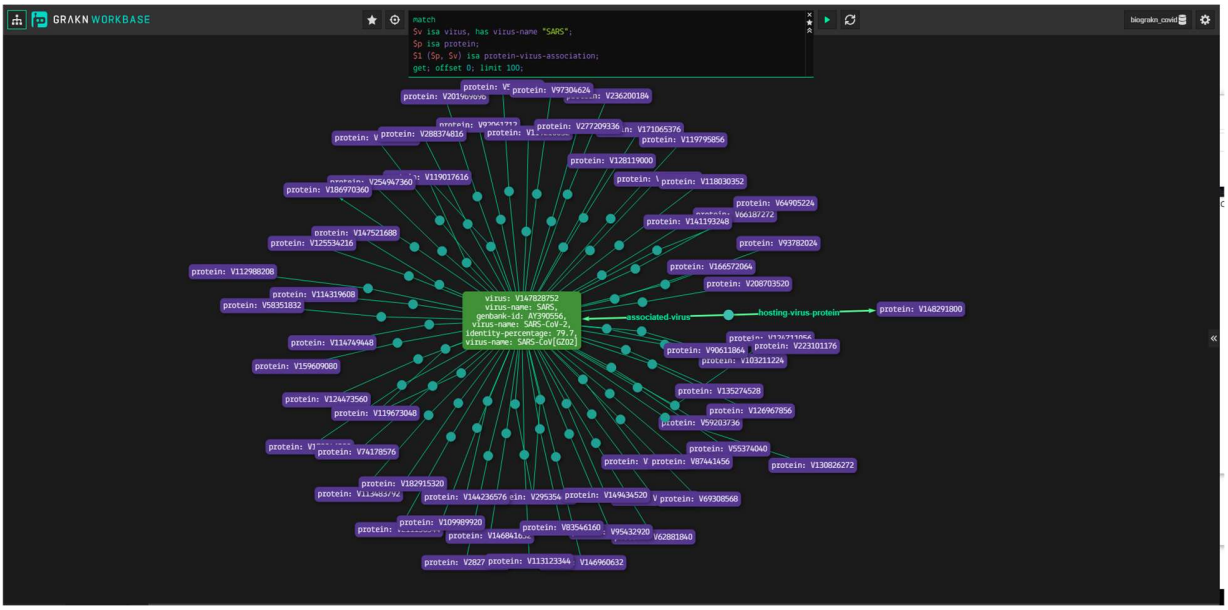

Figure 6 Query 3: “Get all proteins associated with the virus named “SARS”

The last two queries are essentially a combination of the first two and then a generalized form to identify any other useful relations.

Query 4: “Get all genes that encode proteins and their respective encoded proteins that are associated with the virus named “SARS””

match

$v isa virus, has virus-name "SARS";

$g isa gene;

$1 ($g, $v) isa gene-virus-association;

$p isa protein;

$2 (encoding-gene: $g, encoded-protein: $p) isa gene-protein-encoding;

get; offset 0; limit 100;

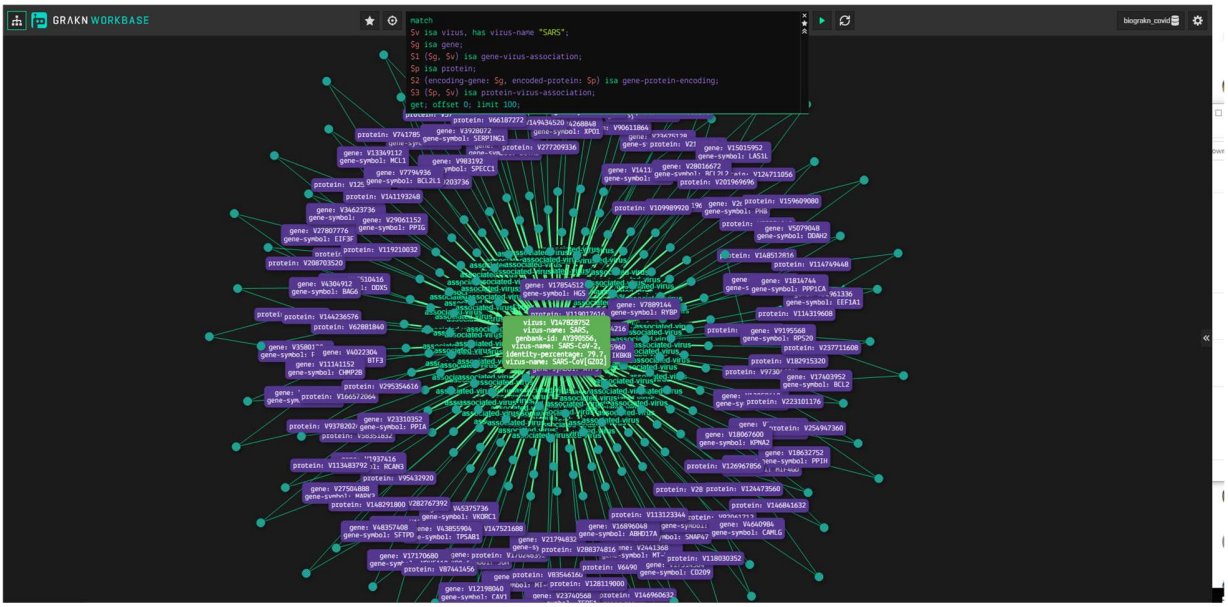

Figure 7 Query 4: Get all genes that encode proteins and their respective encoded proteins that are associated with the virus named “SARS”

Query 5: “Get all genes that encode proteins and their respective encoded proteins that are associated with any coronavirus”

match

$v isa virus;

$g isa gene;

$1 ($g, $v) isa gene-virus-association;

$p isa protein;

$2 (encoding-gene: $g, encoded-protein: $p) isa gene-protein-encoding;

$3 ($p, $v) isa protein-virus-association;

get; offset 0; limit 100;

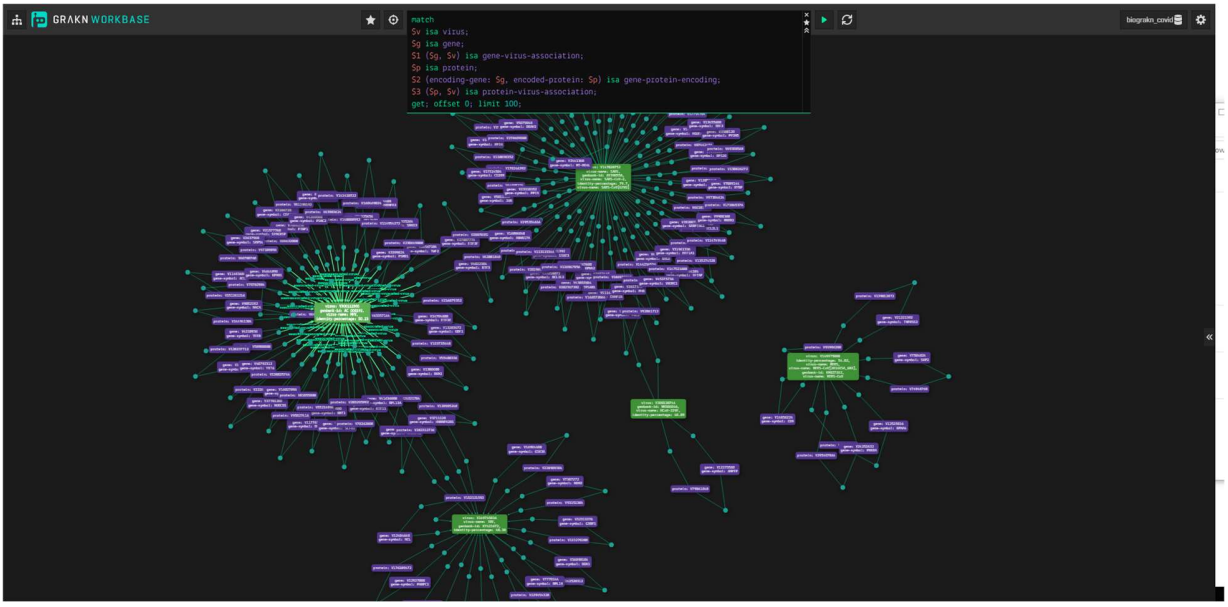

Figure 8 Query 5: “Get all genes that encode proteins and their respective encoded proteins that are associated with any coronavirus”

CHAPTER 5- CONCLUSIONS AND REMARKS

From the aforementioned, it is obvious that Knowledge Graphs can empower most businesses. All businesses that collect and store data about their customers can create a unified profile of their interactions and the relations between them. Furthermore, KGs can reveal hidden relationships between people and products which will help drive business performance. They can be a powerful tool for any sales, customer service or marketing team, by helping them understand the importance of relationships. Besides, businesses can use KGs to uncover new selling opportunities. Marketers can easily visualize the relationships between customers, products and services. Especially when information is pooled from different departments in the same organization, relationships between these entities can be easily addressed.

Furthermore, Knowledge Graphs can help businesses defend against fraud. Financial and legal departments use KGs to uncover patterns between accounts, thus saving time, money and manpower from being involved in a lengthy process. This ability to illuminate fraud relationships is also helpful for companies to manage compliance requirements. With the use of KGs companies can organize customer and account data and eliminate surprises or unwanted headaches for compliance teams. Finally, Knowledge Graphs can precisely adjust the outreaches and reduce expenses (ex. Identify customers who look that they are going to accept an offer, but they wouldn't). This is going to create efficiencies and returns in marketing budgets and KPIs.

In this paper, we examine BioGrakn Covid, a graph-based semantic database that takes advantage of the power of Knowledge Graphs and machine reasoning, to solve problems in the domain of biomedical science and aid Covid-19 related research. A key step is the definition of an ontology, which facilitates the modeling of complex datasets and guarantees information consistency. According to our results, KG are tools that can help by optimizing the complex data management process, by both saving research time, and by better understanding the information provided and the relations within through interactive visualizations and provide useful insights. It should be mentioned that this study did not explore populating a Knowledge Graph by using Text Mining techniques, or by querying Wikidata or other open Knowledge Graphs, not the rest of analytical applications such as Machine Learning or Deep Learning algorithms.

References

Befa, M. & Kontopoulos, E. & Bassiliades, N. & Berberidis, C. & Vahavas, I. (2010), Deploying a Semantically-Enabled Content Management System in a State University. 257-264. 10.1007/978-3-642-15172-9_24.

Koukaras, P., Berberidis C., & Tjortjis C. (2020, August), A Semi-supervised Learning Approach for Complex Information Networks. In Proc. 3rd Int'l conf. Intelligent Data Communication Technologies and Internet of Things (ICICI 2020), Springer Lecture Notes on Data Engineering and Communications Technologies pp. 1-13.

Arenas, M., Grau, B. C., Kharlamov, E., Marciuška, Š., & Zheleznyakov, D. (2016). Faceted search over RDF-based Knowledge Graphs. *Journal of Web Semantics, 37*, pp. 55-74.

Berven, A., Christensen, O., Moldekev, S., & Opdahl, A. (2019). *News Hunter: Building and Mining Knowledge Graphs for Newsroom Systems.* Conference Paper - University of Bergen .

Choudhury, S., Agarwal, K., Purohit, S., Zhang, B., Pirrung, M., Smith, W., & Thomas, M. (2017). Nous: Construction and querying of dynamic Knowledge Graphs. *IEEE 33rd International Conference on Data Engineering (ICDE)*, pp. 1563-1565.

Dwivedi, P. (2020, 05 29). *Creating Knowledge Graphs from Resumes and Traversing them*. Retrieved 09 22, 2020, from https://towardsdatascience.com/creating-knowledge-graphs-from-resumes-and-traver-56016426f4fb

Guo, S., Wang, Q., Wang, L., & Guo, L. (2017). Jointly Embedding Knowledge Graphs and Logical Rules. *National Natural Science Foundation of China* .

Hamilton, W., Bajaj, P., Zitnik, M., Jurafsky, D., & Leskovec, J. (2018). Embedding logical queries on Knowledge Graphs. *Advances in neural information processing systems*, (pp. 2026-2037).

He, S., Liu, K., Ji, G., & Zhao, J. (2015). Learning to represent Knowledge Graphs with gaussian embedding. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* , pp. 623-632.

Heck, L., Hakkani, D., & Tur, G. (2013). Leveraging Knowledge Graphs for Web-Scale Unsupervised Semantic Parsing. *Microsoft Research* .

Huang, Z., Yang, J., van Harmelen, F., & Hu, Q. (2017). Constructing Knowledge Graphs of depression. *International Conference on Health Information Science* (pp. 149-161). Springer, Cham.

Jayaram, N., Khan, A., Li, C., Yan, X., & Elmasri, R. (2015). Querying Knowledge Graphs by example entity tuples. *IEEE Transactions on Knowledge and Data Engineering, 27(10)*, pp. 2797-2811.

Kazemi, S. M., & Poole, D. (2018). Simple embedding for link prediction in Knowledge Graphs. *Advances in neural information processing systems* , pp. 4284-4295.

Kliegr, T., & Zamazal, O. (2016). LHD 2.0: A Text Mining Approach to Typing Entities In Knowledge Graphs . *Journal of Web Semantics* .

Krompaß, D., Baier, S., & Tresp, V. (2015). Type-constrained representation learning in Knowledge Graphs. In International semantic web conference. *International semantic web conference- Springer, Cham.*, pp. 640-655.

Lee, C. W., Fang, W., Yeh, C. K., & Frank Wang, Y. C. (2018). Multi-label zero-shot learning with structured Knowledge Graphs. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1576-1585.

Nickel, M., Murphy, V., Tresp, V., & Gabrilovich, E. (2016, 01). A review of Relational Machine Learning for Knowledge Graphs. *Proceedings of the IEEE, 104(1)*, pp. 11-33.

Pan, J. Z., Vetere, G., Gomez-Perez, J. M., & Wu, H. (2017). *Exploiting linked data and Knowledge Graphs in large organisations.* Heidelberg: Springer.

Paulheim, H. (2016). Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. *Semantic Web*, pp. 1-23.

Popping, R. (2003). Knowledge Graphs and network text analysis. In *Social Science Information* (pp. 91-106). London: SAGE Publications.

Pujara, J., & Getoor, L. (2008). Building Dynamic Knowledge Graphs.

Pujara, J., & Singh, S. (2018). Mining Knowledge Graphs from text. *WSDM*.

Qian, J., Y., L. X., Zhang, C., & Chen, L. (2016). De-anonymizing social networks and inferring private attributes using Knowledge Graphs. *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications* (pp. 1-9). IEEE.

Rospocher, M., van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., & Bogaard, T. (2016). Building Event-Centric Knowledge Graphs from News . *Journal of Web Semantics, 37* , pp. 132-151.

Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., & Sontag, D. (2017). Learning a health knowledge graph from electronic medical records. *Scientific reports, 7(1)*, pp. 1-11.

Trisedia, B. D., Qi, J., & Zhang, R. (2019). Entity Alignment between Knowledge Graphs Using Attribute Embeddings. *The 33rd AAAI Conference on Artificial Intelligence*, pp. 297-304.

Trivedi, P., Maheshwari, G., Dubey, M., & Lehmann, J. (2017). *LC-QuAD: A Corpus for Complex Question Answering over Knowledge Graphs.*

Ulincy, B. (2016). Constructing Knowledge Graphs with Trust.

Voskarides, N., Meij, E., Tsagkias, M., De Rijke, M., & Weerkamp, W. (2015, July). Learning to explain entity relationships in Knowledge Graphs. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* , pp. 564-574.

Wang, X., Wang, D., Xu, C., He, X., Cao, Y., & Chua, T. S. (2019). Explainable reasoning over Knowledge Graphs for recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, (pp. 5329-5336).

Wu, J., Xie, R., Liu, Z., & Sun, M. (2016). Knowledge representation via joint learning of sequential text and Knowledge Graphs. . *arXiv preprint arXiv:1609.07075.*

Xie, R., Liu, Z., & Sun, M. (2016). Representation Learning of Knowledge Graphs with Hierarchical Types. *National Natural Science Foundation of China* .

Ryan A. Rossi and Nesreen K. Ahmed (2015). The Network Data Repository with Interactive Graph Analytics and Visualization. *Association for the Advancement of Artificial Intelligence.*

Lucy Lu Wang, Kyle Lo,, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, Sebastian Kohlmeier1 (2020). Cord-19: The covid-19 open research dataset. *ArXiv, ncbi.nlm.nih.gov*

X Wang, X Song, B Li, Y Guan, J Han (2020). Comprehensive named entity recognition on cord-19 with distant or weak supervision. *arXiv preprint arXiv:2003.12218*

The UniProt Consortium. (2021). UniProt: The universal protein knowledgebase in 2021*, Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D480–D489, https://doi.org/10.1093/nar/gkaa1100*

Sharon L Freshour, Susanna Kiwala, Kelsy C Cotto, Adam C Coffman, Joshua F McMichael, Jonathan J Song, Malachi Griffith, Obi L Griffith, Alex H Wagner, Integration of the Drug–Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts, *Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D1144–D1151, https://doi.org/10.1093/nar/gkaa1084*

Mathias Uhlen, Cheng Zhang, Sunjae Lee, Evelina Sjöstedt, Linn Fagerberg, Gholamreza Bidkhori, Rui Benfeitas, Muhammad Arif, Zhengtao Liu, Fredrik Edfors, Kemal Sanli, Kalle von Feilitzen, Per Oksvold, Emma Lundberg, Sophia Hober, Peter Nilsson, Johanna Mattsson, Jochen M. Schwenk, Hans Brunnström, Bengt Glimelius, Tobias Sjöblom, Per-Henrik Edqvist, Dijana Djureinovic, Patrick Micke, Cecilia Lindskog, Adil Mardinoglu, Fredrik Ponten, (2017). A pathology atlas of the human cancer transcriptome. *science.sciencemag.org, http://www.proteinatlas.org*

Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, Marija Milacic, Corina Duenas Roca, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Guilherme Viteri, Joel Weiser, Guanming Wu, Lincoln Stein, Henning

Hermjakob, Peter D'Eustachio. (2018). *Nucleic Acids Research, Volume 46, Issue D1, Pages D649-D655, https://doi.org/10.1093/nar/gkx1132*

Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, Laura I Furlong. (2019) *The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Research,* doi:10.1093/nar/gkz1021

Halil Kilicoglu, Dongwook Shin, Marcelo Fiszman, Graciela Rosemblat, Thomas C. Rindflesch. (2012), *SemMedDB: a PubMed-scale repository of biomedical semantic predications*. *Bioinformatics, Volume 28, Issue 23, Pages 3158–3160, https://doi.org/10.1093/bioinformatics/bts591*

Koulas, E. (2019). *Defining sovereignty and national interest on cyberspace: national and supranational paradigms.*

Koulas, E., Anthopoulos, M., Grammenou, S., Kaimakamis, C., Kousaris, K., Panavou, F. R., ... & Peristeras, V. (2020). *Misinformation and its stakeholders in Europe: a web-based analysis. arXiv preprint arXiv:2009.09218.*