An IT Recommender System Through Knowledge Graph

Patrick Anderson Matias de Araújo[1], Gian Franco Joel Condori Luna[1]

[1]Instituto de Computação – Universidade Estadual de Campinas (UNICAMP)
Cidade Universitária "Zeferino Vaz" – 13.083-970 – Campinas – SP – Brazil

patrick.araujo@students.unicamp.br, g234826@dac.unicamp.br

Abstract. *The commerce of IT services has risen fast creating a lot of new companies, and this growth become a significant share in the economy of many countries, transforming IT on a whole economic sector. But problems like communication puts a hold on its advancement, as potential clients for IT services struggle to find the best company that serve its needs. Therefore, a recommender system could help these possible buyers find the proper company that offer what the buyer requires. Using knowledge graphs, this research proposal presents a recommender system for this context, for better connect buyers and sellers gaining agility and productivity.*

1. Introduction

Business, trade, and commerce are old activities that has its origins even before capitalism itself. The modern computer, or the electronic computer (ENIAC) was built in 1946 and wouldn't take too much time to business start to explore commercially its benefits. During the XX century humanity has watched the computer evolve and the emergence of big companies like Microsoft, Apple and Google impacting deeply society.

Nowadays there are a whole universe of products and services being provided by IT (Information Technology) companies of many specialties. They also do business with other companies, institutions, and governments. For 2022 is expected that global IT investment reaches $4.5 trillion[1].

As consumers, with so many options, it's easy to get lost. It puts a spotlight in the issue of searching for the right company that provide what you are looking for. This gets worse as projects becomes bigger and more expensive, given the case, many of the choices is made by opinions by someone of trust, such as friends.

Subjectivity could be a problem because who offer the advice doesn't know exactly what you expect, considering your reality, and this means a waste of resources, such as time and money. This reliance on opinions may difficult the best match between client and provider. On the other hand, many big decisions regarding IT on companies aren't necessarily made by specialists on the area, making the definition of requisites ill elaborated.

The goal of this research proposal is to elaborate a recommender system for client and provider regarding IT services to better connect and find the best match between seeker and provider. Initially the proposed system will ask questions to the user, these questions aim to better know the user and its needs and based on these inputs, categorize

[1] Gartner. (2022) "Gartner Forecasts Worldwide IT Spending to Grow 5.1% in 2022". https://www.gartner.com/en/newsroom/press-releases/2022-01-18-gartner-forecasts-worldwide-it-spending-to-grow-five-point-1-percent-in-2022.

the user in a profile and generate a recommendation of technologies as output, that will help to assist the user to achieve what he wants. Using the concepts RDF, OWL, and Turtle a knowledge graph will be made to map the relations between entities, helping provider and seeker matching, and the queries results will be made using SPARQL with the right data.

Currently, lots of new data emerge every single day, becoming a burden for those who wants straightforward information. This is a problem since, just rely on keyword search is not enough to provide what the user wants in a direct manner. A recommender system presents suggestions (output) based on what a user is looking for (input), and this involves several decisions that the system makes, based on a policy. Everyone has already faced a recommender system; it has become essential. There is recommender system for shopping, friends, movies, music, clothes, foods, books, web search and so on.

However, the assembly of a recommender system, urges a big question make the surface: Why it's so hard to recommend? For this very question comes a key issue that is the core of any recommender system: Making the finding of information more assertive. Therefore, there is the need to curate properly the information that is shown to the user. To that end customized recommendations make this process of filtering the information even more curated, many of the popular streaming video services use this artifice, and many factors contribute (geographic information for example) to that.

Although the efforts of the industry, it constitutes a problem the need to grant semantic to data that already exists, and given the case, creating new data with semantics, and this paper will tackle this issue conferring semantics to data and, yet, proving it a specific applicability about a burdensome context in IT commerce. This puts a challenge, once a whole context will need to be transported to a knowledge graph and because it's very specific, it needs to be designing from the ground up, without mentioning the recommender system. This uniqueness is the motivation, because the contribution it provides, enabling the share of knowledge as consequence of the ontology. In the future others system can be deployed on top of the ontology data. The functioning recommender system would be characterization of the validation methodology, returning the proposed information.

2. Background and Related Work

Forged at CERN, the World Wide Web is a hypermedia and information system that cluster documents. The countless documents in this system are interconnected and accessible through the internet. The founder and creator, Tim Berners-Lee (short for Timothy John Berners-Lee) expressed that the current Web "[…] is designed for humans to read, not for computer programs to manipulate meaningfully" (BERNERS-LEE; HENDLER and LASSILA, 2001). Before the current Web the websites were static and without any interaction but clicking in links, this were the Web 1.0, in this version the content showed was always the same, once visited there is no motive to revisit a website and expect something different.

The Web 2.0 changed that perspective, adding bigger interactions in webpages and enabling more social content, that paved the way to social networks like Facebook and Twitter and many other websites that are known to have dynamic content, such as YouTube and Wikipedia that users can contribute and collaborate, this is the current Web.

As a vision of the evolution of the Web, the Semantic Web (Web 3.0) comes along. The Semantic Web extends the current Web (2.0) and add standards and resources to make possible the "Web of data" or the "Web of linked data", rather than the current Web of documents. Linked data is defined as structured data and is integrated with other data becoming interlinked and convenient for semantic queries. For this purpose, "[…] information is given well-defined meaning, better enabling computers and people to work in cooperation" (BERNERS-LEE; HENDLER and LASSILA, 2001).

As stated by World Wide Web Consortium (W3C), the "[…] ultimate goal of the Web of data is to enable computers to do more useful work, and to develop systems that can support trusted interactions over the network" (W3C, 2015) meaning that the "[…] main intent of the Semantic Web is to give machines much better access to information resources so they can be information intermediaries in support of humans" (USCHOLD, 2003), enabling the automation a great deal of tasks with the support of online agents. To that end terms like RDF, OWL, Turtle and, SPARQ comes along as technologies and languages that for the linked data, being the first three used to represent metadata.

Resource Description Framework (RDF) is a model used for description and traffic of interoperable data in the Web and makes it easy to group data even if your schemas are distinct. It uses URI (Uniform Resource Identifier) for name and identify objects connections and allows the sharing, propagation, and union of data of different applications. As result, a direct graph is formed. This direct graph, also known as RDF graph, consist of triples.

Specifically, RDF Schema is an ampliation of the RDF, which provides a vocabulary for data-modelling representation that are used in triple descriptions. Triple, or RDF triple or Semantic triple it's formed by three entities, that it's the atomic data entity in RDF. It encapsules a statement in the shape of subject-predicate-object expressions that resumes semantic data. Turtle (Terse RDF Triple Language), on the other hand, is a syntax that enable to convey data in RDF model. It's also a file format.

OWL (Ontology Web Language) is an ontology language recommended by the W3C and was built upon RDFS. By ontology it "[…] refers to a collection of concepts and axioms that the triples must obey" (REIS e MARTINS, 2022) and constitutes the backbone of a knowledge graph regarding its formal semantics. Knowledge graph is a knowledge model and "[…] describes objects of internet and connections between them" (NOY et al, 2019).

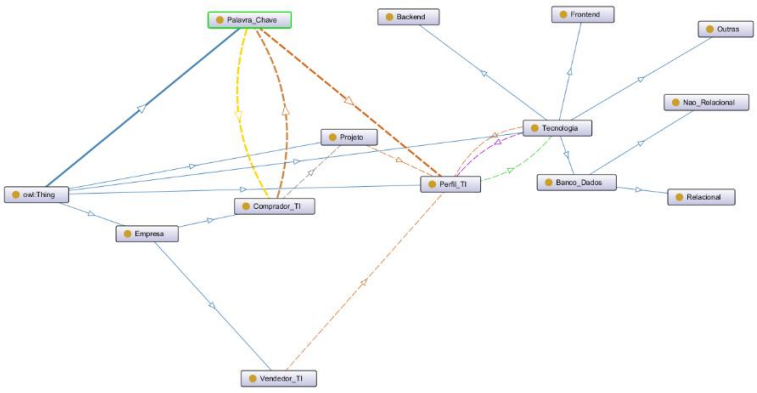

Figure 1. This figure shows an example of Knowledge Graph.

Ontologies were explored in the paper of Khallouki, Abatal and Bahaj (2018) that is about Smart Tourism. The objective of the authors is to design a tourism recommender system for mobile devices with context awareness. Is related that smart tourism is a consequence of smart cities. Therefore, it needs an interconnected, non-static system and that data needs to be traded in real time. A key concept here is semantic plane, that is composed of an ontology and an "[…] intelligent semantic recommender system" (Khallouki; Abatal; Bahaj, 2018) that gather data from multiple sources and provide the required service being the core of the recommender system.

In Awangga et al. (2019) is proposed an ontology from XML data to map knowledge to a specific state program, representing information about family planning. They relate that the Indonesian population is the fourth biggest in the world and the authorities are incapable to provide services to this fast-growing population.

The work of Kontopoulos et al. (2013) question the way sentiment analysis is usually done, not covering all aspects of what is being related from a particular statement. For that, it is proposed the use of ontologies for a more rich and detailed sentiment analysis for Twitter posts. Also in that sense, Ali et al. (2019) used ontologies employing Protégé OWL to "[…] share domain knowledge among different systems" (Ali et al., 2019). The focus of the authors was on make sentiment analysis to help monitor transportation services and traffic control using social networks data posted by the users separating non-relevant content and using machine learning.

Jain, Mehla and Agarwal (2018) proposes a recommender system for the better providing of emergency services and reduce casualties. It's stated that more than half of the land of India is prone to earthquakes, encompassing areas populated. Protégé is used, and ontologies are created for enabling the sharing and reuse of knowledge among different systems, proving interoperability. One of the ontologies is for geographic and geologic information and the other is for emergency services. The goal is for the system offer the better course of action during emergencies, based in similar cases from the past and previous experience.

Focusing on image data, the paper of Kume et al. (2021) relate the building of an ontology about image metadata as data schema. The goal is for an integrated database with many bio-entities for optical and electron microscopy images. In the same line, Jupp et al. (2016) presented the Cellular Microscopy Phenotype Ontology (CMPO) which describes biologic information about the cellular level and phenotypic data. The authors stated that the ontology already is being used to make annotations enabling new data analysis.

The SPARQL Protocol and RDF Query Language was made by W3C as one of the main languages for the semantic web, being a standard for recover data saved in RDF format and kept in databases.

SPARQL and Fuseki server its used for retrieve the queries in the work of Lee et al (2017). They make an ontology, but in this case it's for a tourism recommender system. To that end, in addition to make a travel ontology.

The paper of Danyaro, Jaafar e Liew (2012) utilized SPARQL queries for retrieving triplestores, created and stored in a novel RDF database. It's described the viability of an RDF model for oceanographic and meteorological data. For that the authors

stated that the data in MetOcean (Meteorological and Oceanographic) database was transformed to the RDF domain making this data meaningful and with semantics.

Chen et al. (2020) also make use of SPARQL for queries. Introducing the Protein Ontology Linked Data detailing the PRO RDF data models and its metadata, the main intent is to publicize and integrate knowledge about proteins and its entities on the semantic web using RDF allowing connections with many Linked Open Data enabling more findings in the biological field of knowledge respecting the FAIR (Findability, Accessibility, Interoperability, and Reusability) principle.

In Table 1 there is a resume of the main works described here, the technologies that are used and if a recommender system is proposed.

Table 1. The main works that implements ontologies

Author	Topic / Focus	Recommender System	Technologies Used
Awangga et al. (2019)	Family Planning in Indonesia	No	XML, OWL, and RDF.
Khallouki, Abatal and Bahaj	Smart Tourism	Yes	OWL, RDFS, DAML+OIL, SPARQL.
Danyaro, Jaafar e Liew (2012)	Meteorological and Oceanographic data	No	XML, OWL, RDF, and SPARQL.
Jain, Mehla and Agarwal (2018)	Provide better response about seismic emergences	Yes	OWL, SQL, JSON, and SPARQL.
Chen et al. (2020)	Protein Entities	No	XML, RESTful, RDF, OWL, SPARQL.
Jupp et al. (2016)	Biologic Information about Cellular Level and Phenotypic Data	No	OWL.

Some argue that the semantic web "[…] may or may not come into existence someday" as pointed by Hitzler (2021) because it needs consolidation and mainstream adoption. Some of the biggest players in the tech industry such as Google and Microsoft already have knowledge graphs to "[…] support search and answering questions in search and during conversations" (NOY et al, 2019).

Although the works in this section involves the making of an ontology through a given data, some of them involves a recommender system for providing a service that a person can use, from the specific context. Depending on the case, a new ontology needs to be created, this could be a problem because the need for built a model and gather data from the ground up, but in some papers the reuse of an ontology it's the choice. All the papers offer contribution, providing its perspectives and new ways to use Web Semantics technology.

3. Objectives

The general objective of this project is the creation of a recommendation system for the Brazilian information technology sector through unstructured text. To that end, some specific objectives need to be set:

- Carry out a web scraping of web pages of the Brazilian IT sector;
- Do a text mining of the data obtained by web scraping;
- Create an ontology about the context of IT service buyers and sellers by adding the information obtained in text mining.

4. Materials and Methods

4.1. Web scraping

It is a difficulty that there is no known dataset for the Information Technology sector in the context of Brazil, in consequence of that, an investigation was made in web pages related to this domain. A web search was made and collected a list of URLs from sites of companies in the Brazilian IT sector.

The next step was to perform is a web scraping on them. Currently, the well-known tools for web scraping are BeatifulSoup, Selenium, and Scrapy. BeatifulSoup is a Python package that is very simple to use but also very limited, and it's recommended for basic data extraction. Selenium is a framework created to perform automated web tests but is also used in web scraping. It provides greater interaction with the web page since it allows the friendliest use of its functionalities but consumes considerable resources. Selenium performs better when used for smaller-scale data extraction. Lastly, Scrapy is a Python framework more oriented towards carrying out large-scale tasks, since it has easy creation of threads, and is very efficient when it makes requests asynchronously, which allows a greater speed for extracting data. But Scrapy has is its learning curve, which is a setback, making the learning is higher than the other tools. We will use Scrapy since we have many web pages, and the good use of computational resources and processing time is something very valuable to the scope of this work. Table 2 shows a comparison between Beautiful Soup, Selenium, and Scrapy.

Table 2. Comparison between the most used tools for web scraping

Tool	Easy to use	Asynchronous	Big extraction data	Better Performance
Beautiful Soup	X			
Selenium	X			X
Scrapy		X	X	X

After performing our web page search, were found 195 URLs from the IT sector in Brazil, which we divided into two groups:

- Group 1: These are the informative web pages that companies create to offer their services to clients. They consist mostly of descriptions of the company's objectives, the portfolio of services it provides, and records of works carried out.
- Group 2: These web pages contain a large amount of information prepared by people or companies on a certain domain topic. The web pages store these stories

and display them in the form of lists so that they can be consumed by users, reaching thousands of stories stored on their websites.

For group 1, web scraping was performed, considering that most of these pages contained the main title, a menu of internal links, and paragraphs. To find a particular section within the webpage, we must provide Scrapy with a combination of HTML tags and CSS class names used in the web page. To get the title text we used “h1::text”, to get the menus we use “nav ul > li > a::attr(href)”, and to get the paragraphs we use “p span::text” and also “p::text”. This procedure was carried for each page contained within the internal links menu. However, it was a concern not to navigate in a domain other than that of our main page.

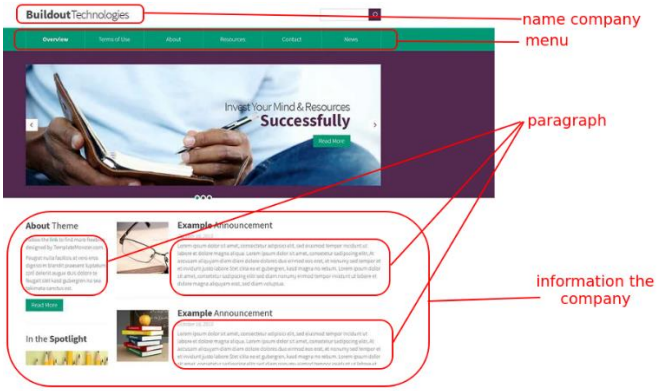

Figure 2. A Web Page and its contents

In group 2, the main page does not display the complete information to be extracted, since it shows a list of stories, but only a small fragment of its content. So, before extracting the information, an interaction with the web page had to be simulated first, allowing to click on each story to view the complete information. A similar interaction was performed to advance through the pager of the web page and to be able to visit all the contained stories. To get the list of internal links, also called menu, was used the combination “.menu-item-has-children:first-child > ul > li > a::attr(href)” and “.menu-topo-categories ul li a::attr(href)” that allowed having the URL of each internal page of the website to visit and to obtain the information in a recursive way. From each story, the title was obtained, with the combination “article h1.entry-title::text” and “article h2.entry-title::text”, and the paragraphs with “article p::text” and “article .entry-content p::text”. The issue about not leave the domain of the main website was applied here.

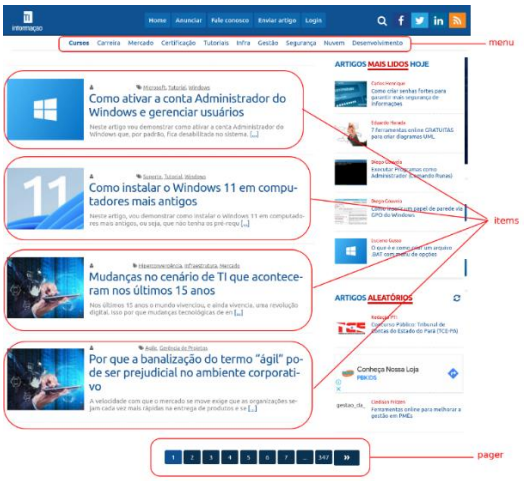

Figure 3. Elements of a Web Page

4.2. Ontology

A lot of factors need to be considered when making an ontology, in this case of IT services, information about companies, and the project that the client wants to get done, are pivotal and need to be in the knowledge graph.

The aim is to develop the ontology using the software Protégé in this current version up to date[2] (5.5.0) and represent the data. As mentioned before, the retrieval of this data will be made using SPARQL. Finally, the recommender system it's going to be developed using the Python programming language using Owlready2's package.

Protégé it's a free and open-source program that runs in Java and it is popular for functions like a knowledge management and an ontology editor. And has as advantage a strong community that provides support and document, a graphical user interface that facilitates its use, the possibility of installing extensions and the support to the "[…] latest OWL 2 Web Ontology Language and RDF specifications from the World Wide Web Consortium" (Protégé, 2022). It has already an ontology semantic reasoner that infer if an ontology follows logical consequences from a set of asserted facts or axioms.

Python it's a very known and popular programming language and will be the chosen language for the implementation of the proposed recommender system. The choice is based on the better applicability of the method, given the Owlready2 package only works in the aforementioned programming language.

Owlready2's package for ontology-oriented programming in Python. "It can load, modify, save ontologies, and it supports reasoning via HermiT" (Owlready2, 2022). It's possible to handle ontology instances, classes, and annotations like they were Python objects. The package is available under the GNU LGPL licence v3.

The recommender system, in a first moment, will make some questions to the user and collect every useful information for better acquaintance with the user and what is desired for the project that this person wants to achieve. Then a categorization will be made based on previous experience that is stored in the ontology, based on the user profile and finally a suggestion of technologies will be generate.

4.3. Activities

To pursue the proposed and reach the objectives, its necessary to follow methodological procedures (activities):

I. Make a literature review;
II. Investigate the context about IT service and review for web pages about the context;
III. Web Scraping in the reviewed pages;
IV. Explore clustering algorithms;
V. Study, create and implement an ontology in the domain of IT services, checking its consistency;
VI. Analysis of the results;
VII. Integration and testing of the Knowledge Graph with the recommender system using the created queries;

[2] October2nd, 2022.

VIII. Write reports and scientific articles about the implementation and its results.

A simplification of the proposed recommender system can be seen in the following figure.

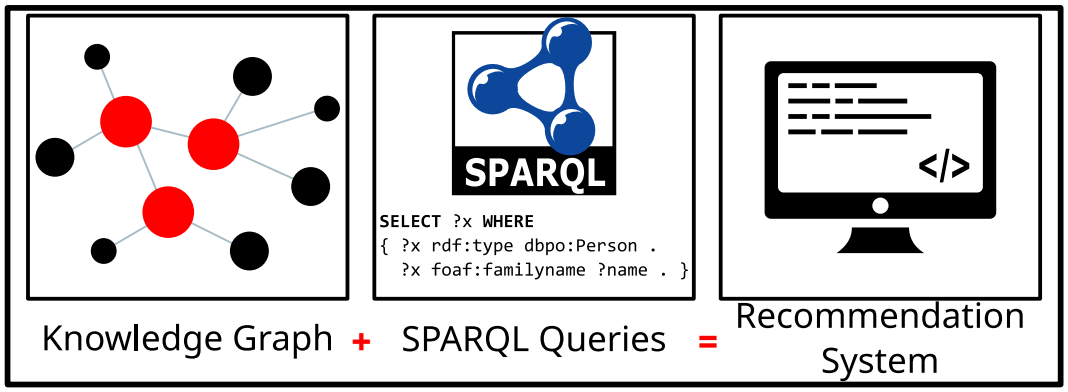

Figure 4. This figure exemplifies the proposed recommender system that consist of a Knowledge Graph and SPARQL Queries.

5. Timeline

Based on the methodological procedures presented on the Materials and Methods section, this the Table 3 is presented the timeline for such activities.

Table 2. Timeline about the foreseen activities

Year	2022				2023											
Month / Activity	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
I	X	X	X	X												
II	X	X	X													
III		X	X													
IV		X	X	X	X	X										
V		X	X	X	X	X	X									
VI					X	X	X	X	X							
VII									X	X	X	X	X			
VIII											X	X	X	X	X	X

References

Ali, Farman, et al. (2019) “Transportation Sentiment Analysis Using Word Embedding and Ontology-Based Topic Modeling”. In: *Knowledge-Based Systems*, vol. 174, page 27–42. https://doi.org/10.1016/j.knosys.2019.02.033.

Awangga, Rolly Maulana, et al. (2019) “Ontology design based on data family planning field officer using OWL and RDF”. In: *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 17, page 161, https://doi.org/10.12928/telkomnika.v17i1.9237.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001) “The Semantic Web. A New Form of Web Content That Is Meaningful to Computers Will Unleash a Revolution of New Possibilities”. In: *Scientific American*, 284, pages 1-5.

Chen, C., Huang, H., Ross, K.E. et al. (2020) Protein ontology on the semantic web for knowledge discovery. In: *Sci Data* 7, 337. https://doi.org/10.1038/s41597-020-00679-9.

Dos Reis, Julio Cesar; Martins, Tulio Brandão Soares. (2022) “Handling Multi-chapter Inconsistencies in DBpedia Evolution”. In: *SIMPÓSIO BRASILEIRO DE BANCO DE DADOS (SBBD)*, 37, 2022, Búzios. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, pages 128-137. ISSN 2763-8979. https://doi.org/10.5753/sbbd.2022.224307.

Hitzler, Pascal. (2021) “A Review of the Semantic Web Field”. Communications of the ACM, vol. 64, pages 76–83. https://doi.org/10.1145/3397512.

Introduction — Owlready2 0.36 documentation. (2022). https://owlready2.readthedocs.io/en/v0.37/intro.html.

Jain, S., Mehla, S., Agarwal, A.G. (2019). An Ontology Based Earthquake Recommendation System. In: Luhach, A., Singh, D., Hsiung, PA., Hawari, K., Lingras, P., Singh, P. (eds) *Advanced Informatics for Computing Research*. ICAICR 2018. Communications in Computer and Information Science, vol 955. Springer, Singapore. https://doi.org/10.1007/978-981-13-3140-4_30

Jupp, Simon, et al. (2016) “The Cellular Microscopy Phenotype Ontology”. In: *Journal of Biomedical Semantics*, vol. 7, page 28. https://doi.org/10.1186/s13326-016-0074-0.

K. U. Danyaro, J. Jaafar and M. S. Liew; (2012) “An RDF model for meteorological and oceanographic information systems”. In: *International Conference on Computer & Information Science (ICCIS)*, 2012, page. 480-484, https://doi.org /10.1109/ICCISci.2012.6297293.

Kontopoulos, Efstratios, et al. (2013) “Ontology-Based Sentiment Analysis of Twitter Posts”. In: *Expert Systems with Applications*, vol. 40, pages 4065–74. https://doi.org/10.1016/j.eswa.2013.01.001.

Kume, S., Masuya, H., Kataoka, Y., & Kobayashi, N. (2016). Development of an Ontology for an Integrated Image Analysis Platform to enable Global Sharing of Microscopy Imaging Data. In: *ISWC2016 The 15th International Semantic Web Conference*. https://doi.org/10.48550/arXiv.2110.10407.

Michael Uschold (2001). Where are the semantics in the semantic web. In: *AI Magazine*, 24, 2003.

Noy, Natasha, et al. (2019) “Industry-Scale Knowledge Graphs: Lessons and Challenges”. In: *Communications of the ACM*, vol. 62, pages 36–43. https://doi.org/10.1145/3331166.

Protégé. (2022). https://protege.stanford.edu/.

Semantic Web - W3C. (2022). https://www.w3.org/standards/semanticweb/.