IN3067/INM713 Semantic Web Technologies and Knowledge Graphs

Introduction. Becoming a Knowledge Scientist

Ernesto Jiménez-Ruiz

Academic course: 2023-2024
Updated: January 29, 2024

Contents

1 Introduction

In this document we introduce the contents of the module and the motivation of including semantic technologies as part of the curriculum. The provided literature in this document can be used by the interested reader to go in depth in some of the areas that the module covers, nevertheless, the recommended reading for the module and for each of the weeks will be provided in the moodle page.

Semantic Web technologies and Knowledge Graphs form the basis of the current development of the Web. At the same time, they have gained significant attention in industry to address use-cases that require exploiting and exchanging heterogeneous and large-scale collections of data. For example, adding semantic understanding to tabular data will be very valuable for data integration, data cleaning, data mining, machine learning and knowledge discovery tasks.

In this module we aim at training the new generation of data scientists with the necessary skills in semantic technologies to meet today's demands from industry, where data (and knowledge) scientists are expected to semantically orchestrate diverse types of data sources.

This module gives a practical combination of web-technology, data management technology, knowledge representation and artificial intelligence. More specifically, we aim at covering the following points:[1]

Semantic Web. The Semantic Web was originally envisioned by Tim Berners-Lee as an extension of the World Wide Web to make Internet data machine-readable. Google's Knowledge Graphs is a prominent example, but the Semantic Web technologies are currently being applied in other environments not necessarily on Web data.

Knowledge Graphs and RDF Graphs. Although they are getting increased popularity, knowledge graphs are not new. Knowledge graphs have their roots in the areas of knowledge representation and Semantic Web. In this module we focus on ontology-layered RDF-based knowledge graphs following World Wide Web Consortium (W3C) standards.

SPARQL query language. SPARQL is the standard W3C language to query and explore RDF-based knowledge graphs. The syntax is relatively similar to SQL, although no previous SQL knowledge is required.

RDF, RDF Schema (RDFS) and Web Ontology Language (OWL) semantics. RDF, RDFS and OWL are W3C standards to represent knowledge graphs and ontologies. We will provide an overview of the semantics and mathematical foundations behind RDFS and OWL, including the OWL profiles: EL, RL and QL.

Inference and Reasoning Engines. Reasoners are an important component of the Semantic Web, as they enable to infer (implicit) knowledge from known axioms, rules and facts. Graphs databases, in particular triplestores, are key engines to perform efficient reasoning over knowledge graphs.

[1] This module reuses content from a related module at the University of Oslo [1], where the module leader contributed in the past.

Ontology and entity alignment. Ontology alignment is key to enable the interoperability in the Semantic Web. Ontology alignment is a also a useful technique in classical data integration, specially when dealing the semantic heterogeneity problem.

Applications to Data Science. Gaining semantic understanding of arbitrary data sources (*e.g.*, tabular data in the form of csv files) will be very valuable for data integration, data cleaning, data mining, machine learning and knowledge discovery tasks. For example, understanding what the data is can help assess what sorts of transformation are appropriate on the data.

Machine learning and knowledge graphs. Machine learning and knowledge graphs (and ontologies) can complement each other. Machine learning approaches, especially neural networks, are typically noise tolerant and can generalize the inference power of ontologies and knowledge graphs. On the other hand, ontologies and knowledge graphs provide sound reasoning capabilities and can provide with enhanced explainability to (black-box) machine learning models.

Extensions. Other points we would like to cover in this module, if time allows, include the extensions RDF* and SPARQL* [28], SHACL language to define constraints [36], and the combination of OWL (RL profile) ontologies and (datalog) rules [25].

2 Semantic Web

Tim Berners-Lee, the inventor of the World Wide Web[2] also had the vision of the Semantic Web [10]:

> "I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web—the content, links, and transactions between people and computers. A 'Semantic Web', which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The 'intelligent agents' people have touted for ages will finally materialize."

Th Web can be seen as a distributed network of hypertext pages that can refer to each other via URLs.[3] The Semantic Web aims at going beyond this, persuading a Web of Data, where each individual data element has its own URI (see Figure 1). The Semantic Web also aims for a (distributed) Web of Data with clear semantics (*i.e.*, well defined meaning) so that it can be processed and orchestrated by smart agents:

[2] First announcement: `https://www.w3.org/People/Berners-Lee/1991/08/art-6484.txt`

[3] **IRI** (Internationalized Resource Identifier), compared to URI, its characters are Unicode which means including Chinese, Japanese and Koreans, etc. **URI** (Uniform Resource Identifier), includes URL and URN. **URL** (Uniform Resource Locator), usually using specific protocols to locate resources within the World Wide Web. **URN** (Uniform Resource Name) uses the *urn:* scheme and it is adopted by the ISBN system.

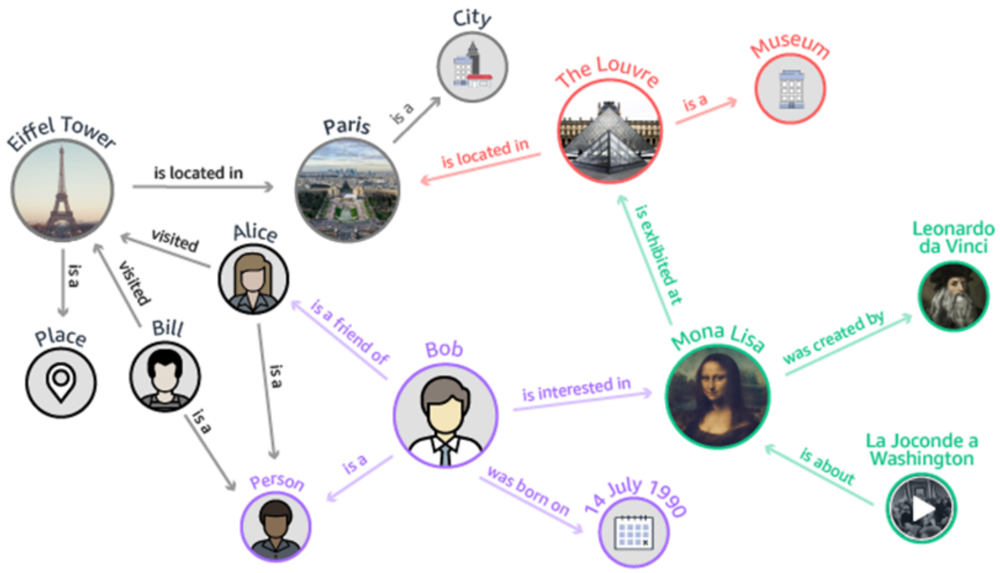


Figure 1: Example of Web of Data (or Knowledge Graph). Extracted from [4].

- Data is published in a machine-readable format.
- Different information sources can be linked.
- Data is enriched with machine-interpretable meaning (domain knowledge) so that (smart) agents can draw conclusions from the available information (see Section 3).

For example, if one would like to know how to get to the cinemas in London that show a comedy film, this will imply more than one Web search to first find a possibly incomplete list of cinemas, check if they are screening comedies and then find out how to get to them. All the necessary information is in principle publicly available (*e.g.*, cinemas, films, genres, screening times, addresses), but not in a format and with the semantics to be easily orchestrated by a smart agent. Google already facilitates a number of searches with its own mash-ups, but it does not contain mash-ups for all possible and desired Web searches.

2.1 Semantic Web Technology Stack

Figure 2 shows the Semantic Web technology stack with the main W3C standards to support the development of the Semantic Web [16].[4] The creation of standards for frameworks and languages is critical *(i)* to have a broader industry (and academic) agreement, *(ii)* to enable interoperability across organizations and applications; and *(iii)* to avoid vendor lock-in of a particular (exchange) format. Fundamental Semantic Web standards:

- RDF (Resource Description Framework) is a simple language or framework for expressing graph data models in the form of triples (subject, predicate, object).

[4] Semantic Web standards: `https://www.w3.org/2001/sw/wiki/Main_Page`

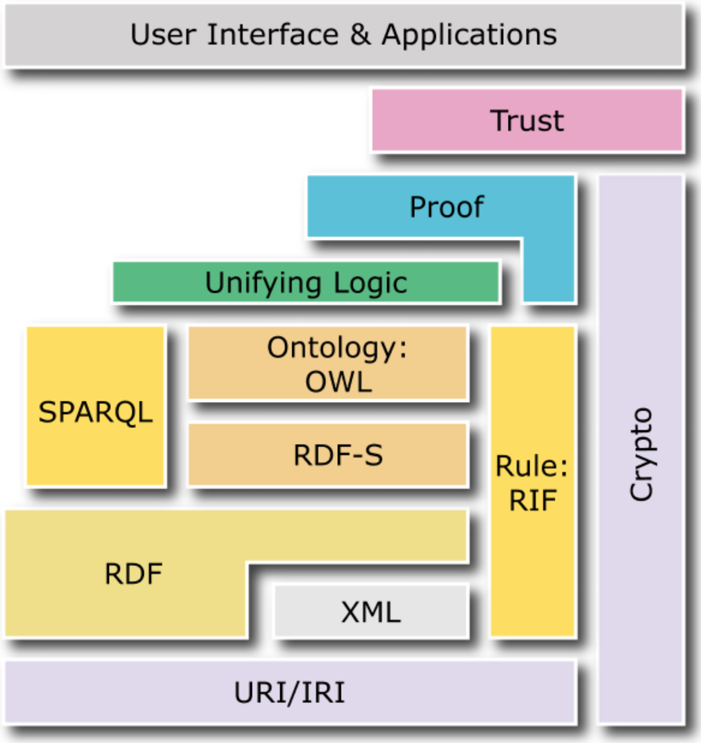


Figure 2: Semantic Web stack and standards [16].

RDF models can be serialized in a variety of syntaxes, *e.g.*, RDF/XML, N3, Turtle, JSON-LD.

- SPARQL is the standard language to query RDF graphs. Unlike SQL, its syntax is designed to match patterns in a graph.
- RDF Schema is a language for describing properties and classes of RDF-based resources, with semantics to represent hierarchies of properties and classes.
- OWL (Web Ontology Language) and its revisions OWL 2 provide a richer language with an extended vocabulary [24]. The formal underpinning of OWL and OWL 2 is based on formal logic [7]. The key advantage of using logic over alternative representation mechanisms is that logic provides an unambiguous meaning to ontologies and enables the use of reasoning engines.

If the information of the example about cinemas screening comedies would be available as an RDF data model, each resource would have an associated unique URI, and RDFS and OWL would provide the semantics of the RDF data model (see Table 1). According to the example triples in Table 1, a smart agent could easily figure out that `http://cinemas/1` is a candidate as it is located in London and it is screening a comedy film. Note that `http://cinemas/2`, assuming no more information is known, would not be a candidate as it is screening a film that is not a comedy.

Table 1: Example RDF triples. rdf:, rdfs: and owl: are the namespaces of the reserved URIs used in the RDF, RDFS and OWL vocabularies, respectively.

subject	predicate	object
http://movies/1	rdf:type	http://movies/Movie
http://movies/1	http://movies/has_genre	http://genres/parody
http://genres/parody	rdf:type	http://movies/Comedy
http://movies/Comedy	rdfs:subClassOf	http://movies/Genre
http://cinemas/1	rdf:type	http://cinemas/Cinema
http://cinemas/1	http://cinemas/screening	http://movies/1
http://cinemas/2	http://cinemas/screening	http://movies/2
http://cinemas/1	http://cinemas/location	http://places/london
http://places/london	rdf:type	http://places/City
http://cinemas/1	http://cinemas/address	"London W1T 1BX"
http://movies/2	http://movies/has_genre	http://genres/zombies
http://genres/zombies	rdf:type	http://genres/Horror
http://genres/Comedy	owl:disjointWith	http://genres/Horror

2.2 From Web of Data to Graph(s) of Knowledge

The notion of distributed Web of Data applied to the whole Web was quite challenging, but we are getting closer to the vision of Tim Berners-Lee. There are increasing efforts in publishing RDF data, *e.g.*, Wikidata [54], DBPedia [6] (Semantic Web version of Wikipedia), the Linked Open Data Cloud[5] includes a variety of datasets, Bio2RDF[6] [9]. Google's Knowledge Graph [50] is also a prominent example. This knowledge graph is used by Google to enhance its search engine results with knowledge gathered from different data sources.

The notion of Web of Data, however, has evolved to a more generic concept: *Graph(s) of Knowledge*. Although this concept is not completely new [27] (see Section 3), the availability of (mature) Semantic Web technology around knowledge graphs is indeed relatively recent. Google has pushed forward the concept of Graph(s) of Knowledge, especially within industry, and nowadays there is an increasing number of companies that have structured their enterprise data as a knowledge graph to drive their products and make them more "intelligent" [44]. Thus, Semantic Web technologies and knowledge graphs can be used not only to *(i)* identify and integrate disparate resources in the Web (*e.g.*, transport, cinemas, films); but also, for example, to *(ii)* integrate data across an organisation (*e.g.*, multiple data sources and departments), *(iii)* combine life science data from genetic, pharmaceutical, patient databases; *(iv)* cross-reference disparate digital libraries.

Exposing data sources as knowledge graphs. Exposing data sources as semantic data typically involves a transformation process or mapping between the original data to a knowledge graph (see Figure 3). Depending on the nature of the data source, the knowledge graph can be materialized or a virtual knowledge graph can be created on demand. The later approach is called ontology based data access (OBDA), and it has also an important application in industry (*e.g.*, [35]).

The W3C has recommendations to generate RDF data from relational

[5] https://lod-cloud.net/

[6] https://bio2rdf.org/

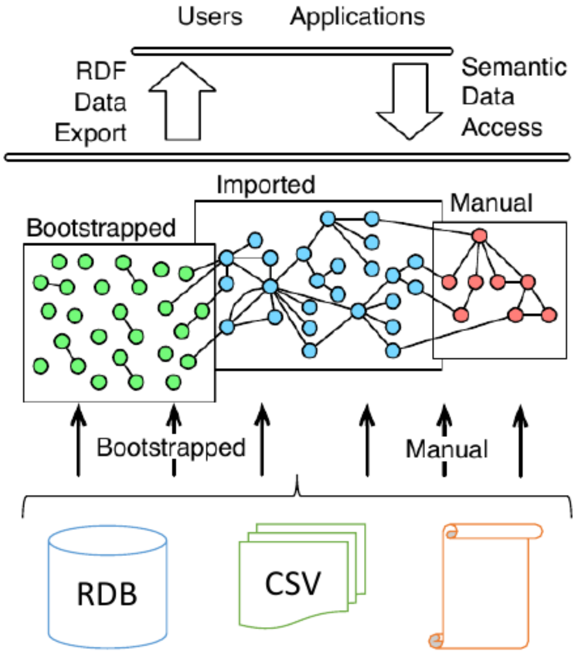


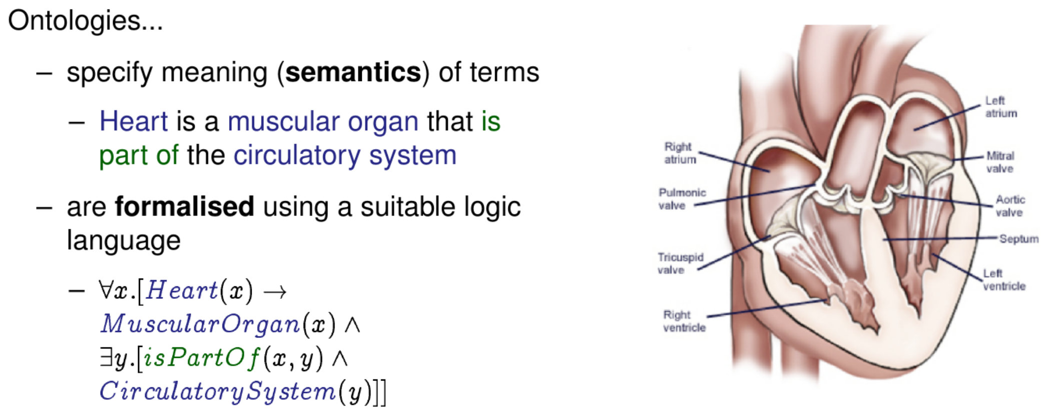
Figure 3: Exposing disparate data sources as semantic data.

databases [5, 18], and CSV files [53] which can lead to an automatically bootstrapped knowledge graph. In addition one could create or reuse a domain knowledge graph as a target for the transformation and perform a semi-automatic transformation. The creation of knowledge graphs from textual resources has been extensively studied in the NLP (Natural Language Processing) community, and specially within the biomedical literature (*e.g.*, [52]).

2.3 Linking Multiple Knowledge Graphs

The Semantic Web does not enforce an agreement to use the same vocabulary in a common ontology or knowledge graph, but to share data that can be integrated with others. This, in practice, leads to the creation of different knowledge graphs for intersecting domains that may agree or not. Nevertheless, the use of resources with clear semantics, instead of just strings, facilitates the interoperability and the detection of potential conflicts and sources of disagreement.

The Semantic Web, unlike relational databases, does not follow the unique naming assumption (UNA), that is, two resources with different URIs can still be the same unless stated the opposite. Ontology alignment or matching [49, 21] is an active research line within the Semantic Web that aims at discovering relationships (aka mappings) among resources with different URIs that represent the same (or related) entities. Ontology alignment builds upon fields like entity resolution and schema matching and is key to enhance interoperability in the Semantic Web. For example, consider the following resources: `http://data1/Paris`, `http://data2/ParisHilton` and `http://data3/Paris_Whitney_Hilton`. All three resources could refer to the same entity, however, `http://data1/Paris` is defined as a `City`, while



Ontologies...

- specify meaning (**semantics**) of terms
  - Heart is a muscular organ that is part of the circulatory system
- are **formalised** using a suitable logic language
  - $\forall x.[Heart(x) \rightarrow MuscularOrgan(x) \wedge \exists y.[isPartOf(x,y) \wedge CirculatorySystem(y)]]$

Figure 4: Example of conceptualization in an ontology (borrowed from [32]).

the other two as `Person`. A smart ontology alignment system, given this knowledge and the fact that things that are a `City` cannot be a `Person` at the same time, will discard a potential equivalence of `http://data1/Paris` to the other two resources. A mapping between the resources `http://data2/ParisHilton` and `http://data3/Paris_Whitney_Hilton` would typically have a high lexical similarity, but other contextual information (*e.g.*, `place of birth`) could increase the confidence of this mapping. Automatic techniques for ontology alignment may bring uncertainty and can potentially predict incorrect mappings. Thus manual revision of (a subset of) the computed mappings is typically required [40].

3 Knowledge Representation and Reasoning

Ontology is a discipline in philosophy that deals with questions about how to represent and categorise entities. In computer science, ontologies play a key role in the development of the Semantic Web and they can be defined as "formal specifications of a shared domain conceptualization" (or as a "abstract symbolic representations of a domain expressed in a formal language") [26, 11]. Figure 4 shows an example of how the semantics of ontology are formalised using a logic language.

Historically, the three main paradigms [43, 48, 47] for the modelling of ontologies have been the network-based paradigm (i.e., semantic networks and other graph-based formalisms), the frame-based paradigm, and the logic-based paradigm.

Semantic networks are a family of formalisms for graphically representing concepts an their relationships. Concepts are represented by nodes, whereas relationships are links connecting nodes. The representation paradigm based on semantic networks has been widely used in artificial intelligence to represent knowledge for expert systems. Originally, as stated by Woods [55] and Brachman [12], semantic networks lacked precise semantics to describe concepts and their relationships, which is crucial to enable automated reasoning (*i.e.*, automatically come to reasonable conclusions) in an unambiguous manner.

KL-ONE and *frames* systems were among the first efforts to formalize semantic networks. The language in KL-ONE [15] provided formal semantics to many of the basic constructs used in semantic networks. Concepts were represented and structured based on the idea of *structured inheritance networks* [13] by using structure-forming operations such as specialization, restriction or differentiation. Furthermore, KL-ONE introduced the notion of *automatic classification*, *i.e.*, the computation of the implicit subsumption relations between concepts.

The concept of *frame* was originally introduced by Minsky [42] as an alternative to logic-oriented approaches in order to provide a more natural way to represent knowledge. The frames paradigm was interpreted by many as simply an alternative syntax to first-order logic [29]. There is no doubt, however, that both frames and the KL-ONE language laid the foundations of modern knowledge representation systems.

3.1 Logic-based representations

As way to provide an unambiguous meaning to semantic networks and frames is to map their constructs to an underlying (symbolic) logic.

Symbolic logic is a subset of the formal logic and can be mainly split within propositional logic and predicate logic. Propositional logic deals with simple facts or declarative propositions (e.g., *'Ernesto is a Lecturer'*) and combination of them using logic connectives (eg. conjunction $\vee$, disjunction $\wedge$, negation $\neg$, implication $\rightarrow$, and biconditional $\leftrightarrow$). Predicate logic is distinguished from propositional logic by the use of predicates (relationships), which can refer to one argument (e.g., $Lecturer(ernesto)$, $University(city)$) or two (e.g., $TeachesIn(ernesto, city)$), and the use of quantifiers ($\forall$, $\exists$) over variables in formulas (e.g., $\forall x.(Lecturer(x) \rightarrow Academic(x))$). Within predicate logic we can distinguish between First-Order Logic (FOL), second-order logic or high-order logic. The difference relies on the type of variable used in quantifications. For example, FOL variables ranges over individuals whereas in second-order logic variables can also range over sets of individuals.

As stated by Hayes [29], the intended meaning of the basic constructs available in semantic networks and frame-based systems could be formalized using FOL. However, the validity problem as well as other central reasoning problems are undecidable in FOL— that is, there is no algorithm that can decide the validity of an arbitrary FOL formula. Gödel and Turing provided valuable insights about the limits of formal logic [51]. Levesque and Brachman [14, 39] emphasized the need for knowledge representation languages that provide a *trade-off* between expressive power and nice computational properties, and they identified a (decidable) fragment of FOL that could express the basic constructs available in semantic networks and frame systems. Originally this subset was called *terminological language* or *concept language*, and eventually evolved into a family of knowledge representation languages called *Description Logics* (DL) [8].[7]

KL-ONE can be seen as the predecessor of current DL-based systems. As in KL-ONE, DL systems make an explicit distinction between the terminological or intensional knowledge (a.k.a. Terminological Box or TBox), which refers to the general knowledge about the domain, and the assertional or extensional knowledge (a.k.a. Assertional Box or ABox), which represents facts about specific individuals. An overview of early DL systems can be found in [20].

[7] In this module DL stands for Description Logics and not Deep Learning.

3.2 The OWL Language and Modern DL Systems

The definition of a standard language for the formal definition of ontologies has been for many years a central research topic for the Semantic Web and Knowledge Representation communities [17, 23].

The Ontology Web Language (OWL) is the World Wide Web Consortium (W3C) standard [24] for ontology modelling. There is currently an extensive range of logic-based algorithmic techniques and infrastructure available for OWL. OWL 2 is the current revision of the OWL Semantic Web language. The formal underpinning of OWL 2 ontologies is based on the description logic $\mathcal{SROIQ}$ (refer to [33] for an extensive description).

Reasoning with OWL ontologies is decidable but tractability may still be a problem in some cases. For this reason the Semantic Web community has defined three lightweight sub-languages or profiles of OWL 2 [37] with very interesting expressiveness and computational properties: *(i)* OWL 2 EL profile is extensively used in biomedical ontologies, *(ii)* OWL 2 RL is key to perform reasoning with (large) Web data, and *(iii)* OWL 2 QL is essential in ontology based data access (OBDA) systems over relational databases. The most important reasoning tasks in each of the profiles can be computed in polynomial time (*i.e.*, they are tractable).

In this module we will use modern DL systems (*e.g.*, OWL 2 reasoners [46]) and triple store reasoners[8] (also referred to as graph databases). The former group can deal with OWL 2 ontologies (and its profiles), while the later typically rely on OWL 2 RL ontologies.

Ontologies and Knowledge Graphs. As described in this section, although Google has relaunched the interest of knowledge graphs in industry, the use of graph models to represent data has been extensively studied in AI. The core idea of knowledge graphs is the enhancement of the graph data model with knowledge. In this module we focus on OWL-layered RDF-based knowledge graphs [31].

Ontologies and Databases. Unlike (relational) databases, ontologies and knowledge graphs *(i)* present a domain model that is independent of the physical schema, *(ii)* include a more user-friendly vocabulary closer to domain experts (*e.g.*, table names in relational databases are not necessarily chosen by domain experts); and *(iii)* provide a flexible model to represent and store semi-structured data, which will enhance the integration of heterogeneous data sources and formats.

4 Motivation: KGs for AI and Data Science

Knowledge graphs organise data from multiple sources, capture information about entities of interest in a given domain or task (like people, places or events), and forge connections between them. In data science and AI, knowledge graphs are commonly used to [45]:

- Facilitate access to and integration of data sources;

[8] `https://www.w3.org/2001/sw/wiki/ToolList`

Table 2: Excerpts of *(a)* a Web table about countries and capitals, *(b)* a real CSV file about broadband data, and *(c)* a custom table with start-ups from Oxford and their foundation year.

(a) Web table

China	Beijing
Indonesia	Jakarta
Congo	Kinshasa
Brazil	
Congo	Brazzaville

(b) CSV file

Virgin	60	London
BT	60	East
BT	40	Scotland
Virgen	40	Wales
Orange	30	West Midlands

(c) Custom table

OST	2017
DeepReason.ai	2018
Oxstem	2011
Oxbotica	2014
DeepMind	2010

- Add context and depth to other, more data-driven AI techniques such as machine learning; and

- Serve as bridges between humans and systems, such as generating human-readable explanations, or, on a bigger scale, enabling intelligent systems for scientists and engineers.

4.1 Semantic Understanding of Tabular Data

Tabular data in the form of CSV files is the common input format in a data analytics pipeline. However, a lack of understanding of the semantic structure and meaning of the content may hinder the data analytics process. Thus, gaining this semantic understanding will be very valuable for data integration, data cleaning, data mining, machine learning and knowledge discovery tasks. For example, understanding what the data is can help assess what sorts of transformation are appropriate on the data.

Tabular data to knowledge graph matching is the process of assigning semantic tags from knowledge graphs to the elements of a table [34]. For example, `Congo` with capital `Brazzaville` (in Table 2a) can be associated to the DBpedia KG entity `http://dbpedia.org/resource/Republic_of_the_Congo`. This task is a challenging problem for various reasons, including but not limited to: *(i)* Lack of metadata or uninformative table and column names, a typical scenario in Web tables and real-world tabular data. *(ii)* Noisiness in the data (*e.g.*, "Virgen" in Table 2b). *(iii)* Knowledge gap, cells without a correspondence to the KG (*e.g.*, relatively recent Oxford start-ups in Table 2c). *(iv)* Ambiguous cells pointing to more than one possible entity (*e.g.*, "Congo" in Table 2a or "Virgin" and "Orange" in Table 2b). *(v)* Missing data (*i.e.*, cells without a value) increasing the effect of the knowledge gap (*e.g.*, capital of "Brazil" in Table 2a). *(vi)* Short labels or acronyms, which typically bring more ambiguity to the matching task (*e.g.*, "BT" in Table 2b).

The result of matching tabular data to a knowledge graph has the potential of providing significant insights about the context of the data and may have an important impact to address the most common data wrangling challenges:

- Data dictionary: basic types and semantic types.

- Data integration from multiple sources

- Entity resolution: duplication and record linkage.

- Format variability: *e.g.*, dates and names.
- Structural variability in the data.
- Identifying and repairing missing data.
- Anomaly detection and repair.
- Metadata/contextual information: (Semantic) data governance.

4.2 Next Decade in AI

Gary Marcus has recently defined the next steps for AI where he highlights the need of richer AI systems, *i.e.*, semantically sound, explainable, and reliable [41]. The new advances in Deep Learning have achieved impressive results to learn patterns on diverse datasets, but they require large amounts of data and they lack explanation of the decisions.

Systems relying on ontologies and knowledge graphs can easily provide explanations about the results of the performed deductive reasoning. These systems, however, rely on rich logical rules and structured knowledge that are hard to build and maintain and may not always be complete. For example, if we have a rule that says $\forall x.\forall y.(A(x) \wedge R(x,y) \wedge B(y) \rightarrow C(x)))$ and we have as data: $A(a)$, $B'(b)$ (being B' similar to B but not quite B) and $R(a,b)$, a system based on deductive reasoning will not be able to imply $C(a)$, however, this could still be possible with machine leaning if there is enough evidence that B and B' are close enough.

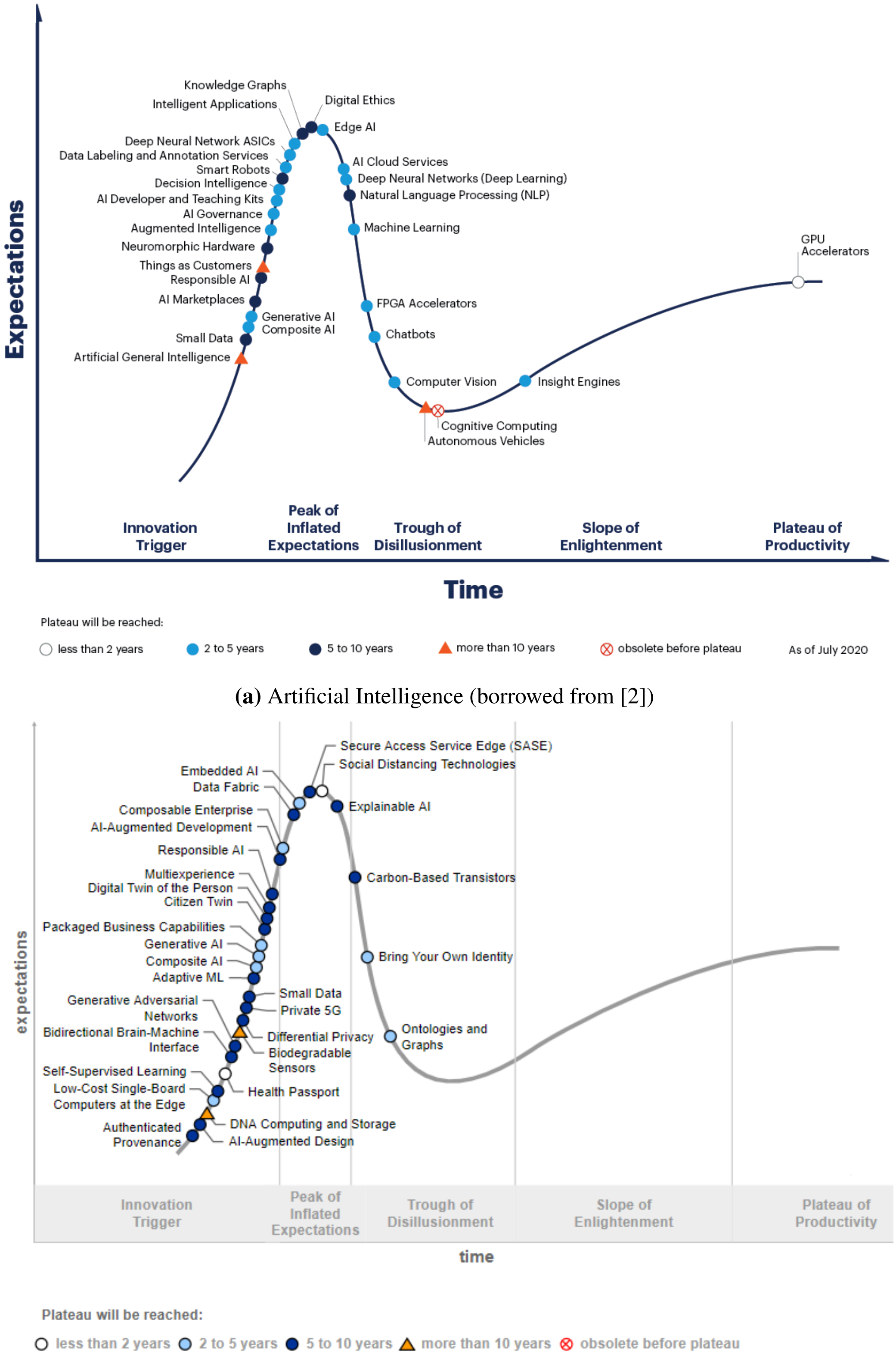
A very promising, although not new, research line is the combinations of connectionist or sub-symbolic systems (*i.e.*, based on neural networks) with symbolic systems (*i.e.*, based on logical representations) to overcome the limitations of both approaches in isolation [30, 38, 19].[9] Ontologies and knowledge graphs can provide structured context that can enhance the learning process and support the explainability. At the same time, models based on neural networks can facilitate the learning of new knowledge. The key is to learn not just patterns in the data, but learning meaning [41] that can be used to extend the symbolic representation.

4.3 Hype Cycle for Emerging Technologies

Figures 5a and 5b shows the technology hype in AI and emerging technologies provided by Gartner.[10] It is very interesting to see Ontologies and Graphs in the "Trough of Disillusionment" (Figure 5b) phase while Knowledge Graphs are still in the "Peak of Inflated Expectations" (Figure 5a). On the one hand, knowledge graphs have recently gained significant attention in both industry and academia; but on the other hand, ontologies and graphs are mature technologies. I would personally place *ontologies and knowledge graphs* closer to the "Slope of Enlightenment" if they are seen as a single technology.

[9] Neural-Symbolic Learning and Reasoning: `http://www.neural-symbolic.org/`

[10] `https://www.gartner.com/en/documents/3887767`



(a) Artificial Intelligence (borrowed from [2])

(b) Emerging Technologies (borrowed from [3])

Figure 5: Gartner's 2020 Hype Cycles

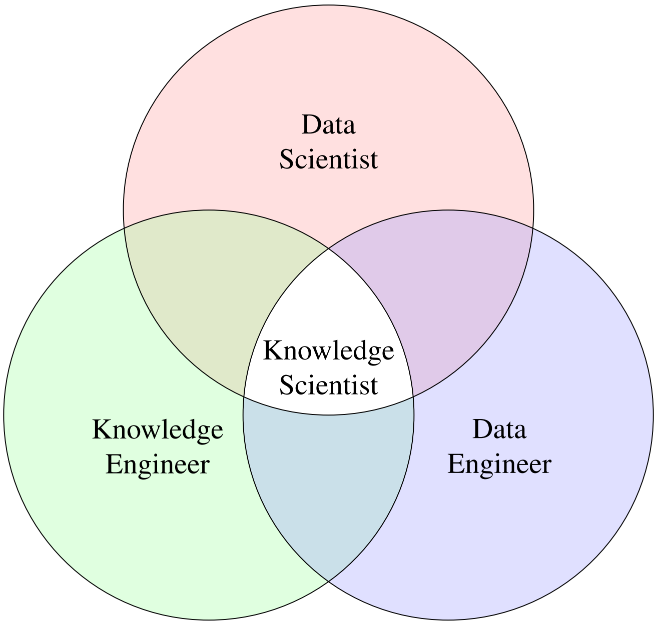


Figure 6: Intersection among the Data Scientist, Knowledge Engineer and Data Engineer professional figures.

5 The Knowledge Scientist

The figure of Knowledge Scientist proposed by Fletcher et al. [22] fits very well to the new demands in data-driven companies. This professional profile is not completely new as it has its roots in the knowledge engineering domain. Data scientist are often covering a large set of tasks which intersect with those of Data Engineers and Knowledge Engineers. Data Scientist typically need to perform data wrangling tasks (see Section 4.1) and understand the data and its context, so that the subsequent analytics steps are meaningful. The reliability of the data is in principle shared among the the Data Engineers and Data Scientist, but in many cases the reliability is shifted to the data wrangling and analytics steps. The figure of the Knowledge Scientist can be seen as an intersection of the Data Scientist, Data Engineer and Knowledge Engineer professional profiles (see Figure 6):

- **Data Engineer:** harnesses and collects large amounts of data.
- **Data Scientist:** draws value and insights from data.
- **Knowledge Engineer:** encodes domain expertise within a computing system.
- **Knowledge Scientist:** adds context to the data to make it more useful, clean, reliable and ready to be used by downstream analytics or AI-based systems.

The Knowledge Scientist role focus on creating a bridge between the data and the business requirements and questions. The output of a Knowledge Scientist is a data model (*e.g.*, a knowledge graph) that represents how the business users see the world [22]. Knowledge Scientist will also serve as bridge communicators between Data Engineers and Data Scientist. Knowledge Scientists will drive a semantic-lifting of the

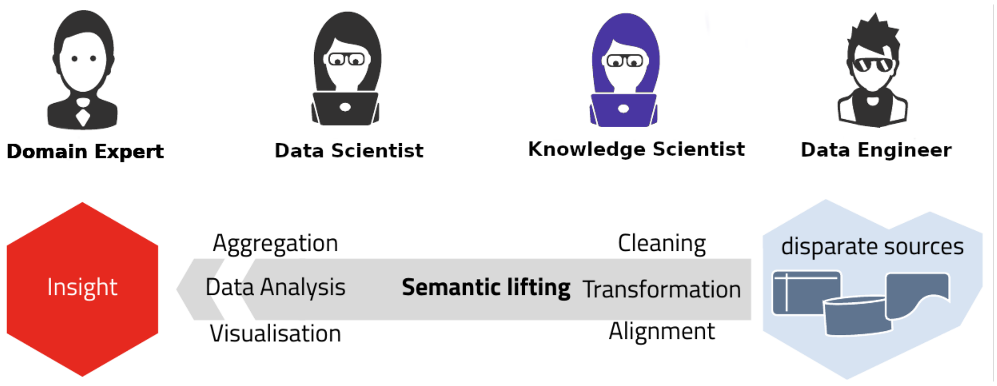


Figure 7: From data to insights (adapted from a slide of Martin Giese, University of Oslo).

data provided by the Data Engineers to facilitate its interpretation by the Data Scientists (see Figure 7). Knowledge Scientists will rely, among others, on the technology and skills we will cover in this module (*e.g.*, data modelling, data integration, knowledge representation, ontology engineering).

In the literature, the Analytics Engineer or Data Steward roles intersect with the notion of Knowledge Scientist, but without a clear emphasis on the need of knowledge representation to semantically encode the domain and the business needs.

References

[1] INF4580 – Semantic technologies (University of Oslo), 2018. `https://www.uio.no/studier/emner/matnat/ifi/INF4580/index.html`.

[2] Gartner's 2020 Hype Cycle for AI, 2020. `https://tinyurl.com/gartner-hype-ai`.

[3] Gartner's 2020 Hype Cycle for Emerging Technologies, 2020. `https://tinyurl.com/gartner-hype-emerging`.

[4] Knowledge Graphs on AWS, Accessed January 2021. `https://aws.amazon.com/neptune/knowledge-graphs-on-aws/`.

[5] Marcelo Arenas, Alexandre Bertails, Eric Prud'hommeaux, and Juan Sequeda. A Direct Mapping of Relational Data to RDF, 2012. W3C Recommendation: `http://www.w3.org/TR/rdb-direct-mapping/`.

[6] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. DBpedia: A Nucleus for a Web of Open Data. In *6th International Semantic Web Conference*, pages 722–735, 2007.

[7] Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[8] Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[9] François Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, and Jean Morissette. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Informatics*, 41(5):706–716, 2008.

[10] Tim Berners-Lee and Mark Fischetti. *Weaving the web - the original design and ultimate destiny of the World Wide Web by its inventor*. HarperBusiness, 2000.

[11] Pim Borst, Hans Akkermans, and Jan Top. Engineering ontologies. *Int. J. Hum.-Comput. Stud.*, 46(2-3):365–406, 1997.

[12] Ronald J. Brachman. What's in a concept: structural foundations for semantic networks. *International Journal of Man-Machine Studies*, 9(2):127 – 152, 1977.

[13] Ronald J. Brachman. On the epistemological status of semantic networks. In N. V. Findler, editor, *Associative networks: Representation and use of knowledge by computers*. New York: Academic, 1979.

[14] Ronald J. Brachman and Hector J. Levesque. The tractability of subsumption in frame-based description languages. In *AAAI*, pages 34–37, 1984.

[15] Ronald J. Brachman and James G. Schmolze. An overview of the KL-ONE knowledge representation system. *Cognitive Science*, 9(2):171 – 216, 1985.

[16] Steve Bratt. Semantic Web, and Other Technologies to Watch, 2007. `https://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/`.

[17] Óscar Corcho and Asunción Gómez-Pérez. A roadmap to ontology specification languages. In *EKAW '00: Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management*, pages 80–96, London, UK, 2000. Springer-Verlag.

[18] Souripriya Das, Seema Sundara, and Richard Cyganiak. R2RML: RDB to RDF Mapping Language, 2012. W3C Recommendation: `https://www.w3.org/TR/r2rml/`.

[19] Artur d'Avila Garcez and Luís C. Lamb. Neurosymbolic AI: the 3rd wave. *CoRR*, abs/2012.05876, 2020.

[20] Jon Doyle. Special issue on implemented knowledge representation and reasoning systems. *SIGART Bull.*, 2(3), 1991.

[21] Jérôme Euzenat and Pavel Shvaiko. *Ontology Matching, Second Edition*. Springer, 2013.

[22] George Fletcher, Paul Groth, and Juan F. Sequeda. Knowledge Scientists: Unlocking the data-driven organization. *CoRR*, abs/2004.07917, 2020.

[23] Asunción Gómez-Pérez and Oscar Corcho. Ontology specification languages for the semantic web. *IEEE Intelligent Systems*, 17(1):54–60, 2002.

[24] Bernardo Cuenca Grau, Ian Horrocks, Boris Motik, Bijan Parsia, Peter F. Patel-Schneider, and Ulrike Sattler. OWL 2: The next step for OWL. *J. Web Semant.*, 6(4):309–322, 2008.

[25] Benjamin N. Grosof, Ian Horrocks, Raphael Volz, and Stefan Decker. Description logic programs: combining logic programs with description logic. In *Proceedings of the Twelfth International World Wide Web Conference (WWW)*, pages 48–57, 2003.

[26] Thomas R. Gruber. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In N. Guarino and R. Poli, editors, *Formal Ontology in Conceptual Analysis and Knowledge Representation*, 1993. Updated Definition: http://tomgruber.org/writing/ontology-definition-2007.htm.

[27] Claudio Gutierrez and Juan F. Sequeda. A Brief History of Knowledge Graph's Main Ideas: A tutorial, 2019. `http://knowledgegraph.today/paper.html`.

[28] Olaf Hartig. Foundations of RDF$\star$ and SPARQL$\star$ (An Alternative Approach to Statement-Level Metadata in RDF). In *Proceedings of the 11th Alberto Mendelzon International Workshop on Foundations of Data Management and the Web*, 2017.

[29] Patrick J. Hayes. The logic of frames. In D. Metzing, editor, *Frame Conceptions and Text Understanding*, pages 46–61. Walter de Gruyter and Co., Berlin, 1979.

[30] Pascal Hitzler, Federico Bianchi, Monireh Ebrahimi, and Md. Kamruzzaman Sarker. Neural-symbolic integration and the semantic web. *Semantic Web*, 11(1):3–11, 2020.

[31] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutiérrez, José Emilio Labra Gayo, Sabrina Kirrane, Sebastian Neumaier, Axel Polleres, Roberto Navigli, Axel-Cyrille Ngonga Ngomo, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan F. Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge Graphs. *CoRR*, abs/2003.02320, 2020.

[32] Ian Horrocks. Ontologies and the Semantic Web: The Story So Far, 2017. Seminars: `http://www.cs.ox.ac.uk/people/ian.horrocks/Seminars/seminars.html`.

[33] Ian Horrocks, Oliver Kutz, and Ulrike Sattler. The even more irresistible $\mathcal{SROIQ}$. In *Proceedings, Tenth International Conference on Principles of Knowledge Representation and Reasoning*, pages 57–67, 2006.

[34] Ernesto Jimenez-Ruiz, Oktie Hassanzadeh, Vasilis Efthymiou, Jiaoyan Chen, and Kavitha Srinivas. SemTab 2019: Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems. In *The Semantic Web: ESWC 2020*. Springer International Publishing, 2020.

[35] Evgeny Kharlamov, Dag Hovland, Martin G. Skjæveland, Dimitris Bilidas, Ernesto Jiménez-Ruiz, Guohui Xiao, Ahmet Soylu, Davide Lanti, Martin Rezk, Dmitriy Zheleznyakov, Martin Giese, Hallstein Lie, Yannis E. Ioannidis, Yannis Kotidis, Manolis Koubarakis, and Arild Waaler. Ontology based data access in statoil. *J. Web Semant.*, 44:3–36, 2017.

[36] Holger Knublauch and Dimitris Kontokostas. Shapes Constraint Language (SHACL), 2017. W3C recommendation: `https://www.w3.org/TR/shacl/`.

[37] Markus Krötzsch. OWL 2 Profiles: An Introduction to Lightweight Ontology Languages. In *Reasoning Web. Semantic Technologies for Advanced Query Answering - 8th International Summer School 2012, Vienna, Austria, September 3-8, 2012. Proceedings*, pages 112–183, 2012.

[38] Luís C. Lamb, Artur S. d'Avila Garcez, Marco Gori, Marcelo O. R. Prates, Pedro H. C. Avelar, and Moshe Y. Vardi. Graph Neural Networks Meet Neural-Symbolic Computing: A Survey and Perspective. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4877–4884, 2020.

[39] Hector J. Levesque and Ronald J. Brachman. Expressiveness and tractability in knowledge representation and reasoning. *Computational Intelligence*, 3:78–93, 1987.

[40] Huanyu Li, Zlatan Dragisic, Daniel Faria, Valentina Ivanova, Ernesto Jiménez-Ruiz, Patrick Lambrix, and Catia Pesquita. User validation in ontology alignment: functional assessment and impact. *Knowl. Eng. Rev.*, 34:e15, 2019.

[41] Gary Marcus. The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. *CoRR*, abs/2002.06177, 2020.

[42] M. Minsky and P. H. Winston. A framework for representing knowledge. In *The Psychology of Computer Vision*, New York, 1975. McGraw Hill.

[43] Daniele Nardi and Ronald J. Brachman. An introduction to Description Logics. In Baader et al. [8], pages 5–44.

[44] Natalya Fridman Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. Industry-scale knowledge graphs: lessons and challenges. *Commun. ACM*, 62(8):36–43, 2019.

[45] Jeff Pan, Elena Simperl, Ernesto Jiménez-Ruiz, and Ian Horrocks. Knowledge Graphs Interest Group, 2021. The Alan Turing Institute: `https://www.turing.ac.uk/research/interest-groups/knowledge-graphs`.

[46] Bijan Parsia, Nicolas Matentzoglu, Rafael S. Gonçalves, Birte Glimm, and Andreas Steigmiller. The OWL reasoner evaluation (ORE) 2015 competition report. *J. Autom. Reason.*, 59(4):455–482, 2017.

[47] Ismael Sanz and Ernesto Jiménez-Ruiz. Ontologías en informática. In A. Alcina, E. Valero, and E. Rambla, editors, *Terminología y Sociedad del conocimiento*, pages 255–286. Peter Lang, Berna, 2009.

[48] Ulrike Sattler, Diego Calvanese, and Ralf Molitor. Relationships with other formalisms. In Baader et al. [8], pages 142–183.

[49] Pavel Shvaiko and Jérôme Euzenat. Ontology matching: State of the art and future challenges. *IEEE Trans. Knowl. Data Eng.*, 25(1):158–176, 2013.

[50] Amit Singhal. Introducing the Knowledge Graph: Things, Not Strings, 2012. `https://blog.google/products/search/introducing-knowledge-graph-things-not/`.

[51] Michael Sipser. *Introduction to the theory of computation*. PWS Publishing Company, 1997.

[52] Irena Spasic, Sophia Ananiadou, John McNaught, and Anand Kumar. Text mining and ontologies in biomedicine: Making sense of raw text. *Briefings Bioinform.*, 6(3):239–251, 2005.

[53] Jeremy Tandy, Ivan Herman, and Gregg Kellogg. Generating RDF from Tabular Data on the Web, 2015. W3C Recommendation: `https://www.w3.org/TR/csv2rdf/`.

[54] Denny Vrandecic and Markus Krötzsch. Wikidata: a free collaborative knowledge base. *Commun. ACM*, 57(10):78–85, 2014.

[55] William A. Woods. What's in a link: Foundations for semantic networks. In D. Bobrow and A. Collins, editors, *Representation and Understanding: Studies in Cognitive Science*. Academic Press, New York, NY, 1975.