

Towards a Definition of Knowledge Graphs

Lisa Ehrlinger and Wolfram Wöb
Institute for Application Oriented Knowledge Processing
Johannes Kepler University Linz, Austria
{lisa.ehrlinger | wolfram.woess}@jku.at

ABSTRACT

Recently, the term *knowledge graph* has been used frequently in research and business, usually in close association with Semantic Web technologies, linked data, large-scale data analytics and cloud computing. Its popularity is clearly influenced by the introduction of Google’s Knowledge Graph in 2012, and since then the term has been widely used without a definition. A large variety of interpretations has hampered the evolution of a common understanding of knowledge graphs. Numerous research papers refer to Google’s Knowledge Graph, although no official documentation about the used methods exists. The prerequisite for widespread academic and commercial adoption of a concept or technology is a common understanding, based ideally on a definition that is free from ambiguity. We tackle this issue by discussing and defining the term knowledge graph, considering its history and diversity in interpretations and use. Our goal is to propose a definition of knowledge graphs that serves as basis for discussions on this topic and contributes to a common vision.

CCS Concepts

•Information systems → Data management systems; Information systems applications;

Keywords

Knowledge Graphs, Knowledge Bases, Ontologies, Knowledge Representation, Semantic Web.

1. INTRODUCTION

Considerable research into knowledge graphs (KGs) has been carried out in recent years, especially in the Semantic Web community, and thus a variety of partially contradicting definitions and descriptions has emerged. The often quoted blog entry by Google [18] basically describes an enhancement of their search engine with semantics. And also Wikipedia, the most comprehensive encyclopedia in the web, does not provide information about knowledge graphs in general, but refers to the implementation by Google without mentioning the existence of other knowledge graphs. Although Wikipedia is no scientific reference source, it contributes to a common understanding through its role as a primary information

source for several prominent knowledge representation applications. Other definitions may lead to the assumption that knowledge graph is a synonym for any graph-based knowledge representation (cf. [12, 16]). We argue that such a definition is not enough for an adequate application of knowledge graphs, since it does not enforce a minimum set of requirements a KG has to fulfill. Thus, even a simple graph-based vocabulary could be published as knowledge graph. In addition, such a definition creates an entrance barrier for people who are unfamiliar with knowledge graphs and want to delve deeper into the topic or aim at building a KG on their own. A clear definition propagates a shared understanding of benefits, improvements, or drawbacks that can be expected if someone builds a knowledge graph. Thus, we provide a discussion of knowledge graphs and motivate a definition to support a common understanding.

This paper is organized as follows: Section 2 presents related state-of-the-art research and existing attempts at defining knowledge graphs. Section 3 provides a short overview of historic and current knowledge graph applications and Section 4 delimits and differentiates the term knowledge graph from similar terms and introduces our definition.

2. SELECTED DEFINITIONS

Knowledge graphs have been in the focus of research since 2012 resulting in a wide variety of published descriptions and definitions. Table 1 lists representative definitions and demonstrates the lack of a common core, a fact that is also indicated by Paulheim [16] in 2015. Paulheim listed in his survey of knowledge graph refinement the minimum set of characteristics that must be present to distinguish knowledge graphs from other knowledge collections (cf. first definition in Table 1), which basically restricts the term to any graph-based knowledge representation. In the online reviewing process of “Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods” [16], Noy¹ agreed that a more precise definition was hard to find at that point. This statement points out the demand for closer investigation and deeper reflection in this area.

Vague descriptions of knowledge graphs were published in the announcement of a special issue on knowledge graphs by the Journal of Web Semantics and by the Semantic Web Company (cf. second and third definitions in Table 1). Both definitions could equally well describe an ontology or – even more generally – any kind of semantic knowledge representa-

¹<http://www.semantic-web-journal.net/content/knowledge-graph-refinement-survey-approaches-and-evaluation-methods> [August, 2016]

Definition	Source
“A knowledge graph (i) mainly describes real world entities and their interrelations, organized in a graph, (ii) defines possible classes and relations of entities in a schema, (iii) allows for potentially interrelating arbitrary entities with each other and (iv) covers various topical domains.”	Paulheim [16]
“Knowledge graphs are large networks of entities, their semantic types, properties, and relationships between entities.”	Journal of Web Semantics [12]
“Knowledge graphs could be envisaged as a network of all kind things which are relevant to a specific domain or to an organization. They are not limited to abstract concepts and relations but can also contain instances of things like documents and datasets.”	Semantic Web Company [3]
“We define a Knowledge Graph as an RDF graph. An RDF graph consists of a set of RDF triples where each RDF triple (s, p, o) is an ordered set of the following RDF terms: a subject $s \in U \cup B$, a predicate $p \in U$, and an object $U \cup B \cup L$. An RDF term is either a URI $u \in U$, a blank node $b \in B$, or a literal $l \in L$.”	Färber et al. [7]
“[...] systems exist, [...], which use a variety of techniques to extract new knowledge, in the form of facts, from the web. These facts are interrelated, and hence, recently this extracted knowledge has been referred to as a knowledge graph.”	Pujara et al. [17]

Table 1: Selected definitions of knowledge graph

tion and do not even enforce a graph structure. In addition, size is highlighted as an essential characteristic, which is reflected by phrases such as “large networks” or “vast networks” [11], while it remains unclear what “large” means in this context. Färber et al. defined a knowledge graph as an Resource Description Framework (RDF) graph and stated that the term KG was coined by Google to describe any graph-based knowledge base (KB) [7]. Although this definition is the only formal one, it contradicts with more general definitions as it explicitly requires the RDF data model. Pujara et al. did not provide a concise definition, but rather described the characteristics of knowledge graphs. Unlike the other definitions, which focus solely on the inner structure of the KG, they highlighted the importance of an automatic extraction system. In the preface of the 13th International Semantic Web Conference Proceedings (2014), the following statement was published:

Significantly, major companies, such as Google, Yahoo, Microsoft, and Facebook, have created their own “knowledge graphs” that power semantic searches and enable smarter processing and delivery of data: The use of these knowledge graphs is now the norm rather than the exception. [14]

Once again, this highlights the demand for a common definition, because it is necessary to define and differentiate KGs from other concepts in order to make valuable and accurate statements about the introduction and dissemination of knowledge graphs. Furthermore, this ISWC statement proclaims the use of knowledge graphs to be the norm in general, instead of restricting the scope, domain, or application area where KGs can be used beneficially and efficiently. Despite its lack of clarity, this statement seems to have inspired many researchers to submit papers about knowledge graphs in the following conference in 2015².

3. KNOWLEDGE GRAPH APPLICATIONS

²<http://iswc2015.semanticweb.org/program/accepted-papers> [August, 2016]

In the 1980s, researchers from the University of Groningen and the University of Twente in the Netherlands initially introduced the term knowledge graph to formally describe their knowledge-based system that integrates knowledge from different sources for representing natural language [10, 15]. The authors proposed KGs with a limited set of relations and focus on qualitative modeling including human interaction, which clearly contrasts with the idea of KGs that has been widely discussed in recent years.

In 2012, Google introduced the *Knowledge Graph* as a semantic enhancement of Google’s search function that does not match strings, but enables searching for “things”, in other words, real-world objects [18]. Although the blog post does not provide any implementation details, it has been cited more than 100 times according to Google Scholar³. Since 2012, the term knowledge graph is also used to describe a family of applications. Frequently mentioned implementations are DBpedia, YAGO (Yet Another Great Ontology), Freebase, Wikidata, Yahoo’s semantic search assistant tool Spark, Google’s Knowledge Vault, Microsoft’s Satori and Facebook’s entity graph [7, 14, 16, 11]. Those applications differ in their characteristics, such as architecture, operational purpose, and technology used, which makes it difficult to find a consensus and to create a definition of knowledge graph. The lowest common denominator of the listed open source applications is their use of Linked Data, whereas hardly any proven information is available about Satori and the entity graph.

In addition, the more specific term *enterprise knowledge graph* is used by a few smaller companies, for example, SindiceTech⁴ and the Semantic Web Company [3]. Both companies seek to describe a similar model that extracts and stores diverse enterprise data in a triple store and analyzes it by using machine learning techniques in order to acquire new knowledge from the data and to reuse it in other applications.

³https://scholar.google.at/scholar?q=Introducing+the+Knowledge+Graph%3A+things%2C+not+strings&btnG=&hl=en&as_sdt=0%2C5 [August, 2016]

⁴<http://www.sindicetech.com/overview.html> [August, 2016]

4. TERMINOLOGICAL ANALYSIS AND DEFINITION

When analyzing current research work that defines or addresses knowledge graphs, two fundamental issues can be identified: (a) Google’s blog entry about their Knowledge Graph is cited as if it provides a proper explanation for constituting a knowledge graph (cf. [17, 19]), and (b) the terms knowledge graph and knowledge base are used interchangeably (cf. [5, 7, 8, 13, 16, 20]). The second problem leads to the misleading assumption that the term knowledge graph is a synonym for knowledge base, which is itself often used as synonym for ontology. An example of such confusion is that both Knowledge Vault and Google’s Knowledge Graph have been called large-scale knowledge base by their respective creators [5]. A further example is YAGO, which is – according to its name – an ontology, but is referred to both as knowledge base (cf. [5, 16]) and as knowledge graph (cf. [8, 21]). Similarly, Yahoo employees [2] do not distinguish clearly between knowledge base, knowledge graph and ontology. They state that they construct their knowledge base by aligning new entities, relations and information with their common ontology. Therefore, incomplete, inconsistent and possibly inaccurate information is turned into a rich, unified, disambiguated knowledge graph. Based on this information, their understanding of a knowledge graph is the cleaned knowledge base that is the population (e.g., instances) of their ontology. In order to distinguish between the terms, they must be clarified explicitly. According to Akerkar and Sajja [1], a knowledge-based system uses artificial intelligence to solve problems, and it consists of two parts: a knowledge base and an inference engine. The knowledge base is a dataset with formal semantics that can contain different kinds of knowledge, for example, rules, facts, axioms, definitions, statements, and primitives [4]. Thus, Knowledge Vault cannot be classified as a true knowledge base, because it extends the idea of a pure semantic store with reasoning capabilities and therefore bears more resemblance to a knowledge-based system.

An ontology is as a formal, explicit specification of a shared conceptualization that is characterized by high semantic expressiveness required for increased complexity [9]. Ontological representations allow semantic modeling of knowledge, and are therefore commonly used as knowledge bases in artificial intelligence (AI) applications, for example, in the context of knowledge-based systems. Application of an ontology as knowledge base facilitates validation of semantic relationships and derivation of conclusions from known facts for inference (i.e., reasoning) [9]. We explicitly emphasize that an ontology does not differ from a knowledge base, although ontologies are sometimes erroneously classified as being at the same level as database schemas [6]. In fact, an ontology consists not only of classes and properties (e.g., `owl:ObjectProperty` and `owl:DatatypeProperty`), but can also hold instances (i.e., the population of the ontology).

On the one hand, size is often mentioned as an essential characteristic of knowledge graphs, therefore a KG could be described as very large ontology. However, other contributors have pointed out that knowledge graphs are somehow superior to ontologies [3] and provide additional features. Thus, the difference between a knowledge graph and an ontology could be interpreted either as a matter of quantity (e.g., a large ontology), or of extended requirements (e.g., a built-in reasoner that allows new knowledge to be derived).

The second interpretation leads to the assumption that a knowledge graph is a knowledge-based system that contains a knowledge base and a reasoning engine. Focusing on existing automatically generated “knowledge graphs”, we can identify a further essential characteristic: collection, extraction, and integration of information from external sources extends a pure knowledge-based system with the concept of integration systems. Most open source applications listed in Section 3 implement the integration aspect with Linked Data.

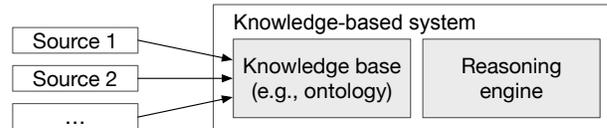


Figure 1: Architecture of a knowledge graph

Figure 1 illustrates the combination of these assumptions, which yields an abstract knowledge graph architecture. Based on this architecture and derived from the terminological analysis, we define a knowledge graph as follows:

A knowledge graph acquires and integrates information into an ontology and applies a reasoner to derive new knowledge.

This definition aligns with the assumption that a knowledge graph is somehow superior and more complex than a knowledge base (e.g., an ontology) because it applies a reasoning engine to generate new knowledge and integrates one or more information sources. Consequently, a manually created knowledge graph that does not support integration aspects is a plain knowledge base or knowledge-based system if it provides reasoning capabilities. The definition does not take the quantity aspect (size) into account, especially with respect to a large ABox of the ontology, since it is not clear what can be considered “large”. Instead, reasoning capabilities are highlighted as an essential characteristic to derive new knowledge and differentiate a KG from knowledge bases.

Furthermore, the question arises what constitutes the difference between the Semantic Web and knowledge graphs. Smaller KGs, for example, enterprise knowledge graphs, can be clearly differentiated from the Semantic Web because of their restricted domain. The goal of large search engines is to crawl and process all available information in the web, which leads to an increased interest in the widespread implementation of semantic technology. Hardly any information is available on the technologies applied in Google’s Knowledge Graph and Microsoft’s Satori, but Yahoo’s Spark and the Knowledge Vault apparently use Semantic Web standards such as RDF. Considering the layers of the Semantic Web, a knowledge graph, in comparison, deploys either exactly the same technology for every layer or a similar one that offers the same features. For example, companies might use proprietary identifiers in place of URIs and JSON-LD⁵ as serialization format substituting XML and RDF. However, these technologies are just examples, and particularly in the syntax layer XML is often replaced with more lightweight and more easily readable formats such as Turtle, N-Triples or N-Quads in the Semantic Web community. In conclusion, the Semantic Web could be interpreted as the most comprehensive knowledge graph, or – conversely – a knowledge

⁵<http://json-ld.org> [August, 2016]

graph that crawls the entire web could be interpreted as self-contained Semantic Web.

5. CONCLUSION

Graph-based knowledge representation has been researched for decades and the term knowledge graph does not constitute a new technology. Rather, it is a buzzword reinvented by Google and adopted by other companies and academia to describe different knowledge representation applications. We have proposed a definition of knowledge graph in order to promote a discussion and a common vision for further work in this area. There are essential differences in the way knowledge representation applications (cf. Section 3) are built, ranging from completely handcrafted knowledge bases to automatically extracted and processed knowledge graphs. Consequently, the term knowledge graph is not suitable for describing all of these applications and should be used more carefully. Several applications need not be called knowledge graphs, because the terms knowledge base and ontology describe them sufficiently and more accurately. Taking into account the diverse applications, a KG bears more resemblance to an abstract framework than to a mathematical structure. Our ongoing research focuses on an in-depth analysis of our definition with respect to existing KG implementations as well as the assessment of data quality in knowledge graphs and their accessed sources.

6. REFERENCES

- [1] R. Akerkar and P. Sajja. *Knowledge-Based Systems*. Jones and Bartlett Publishers, USA, 1st edition, 2009.
- [2] R. Blanco, B. B. Cambazoglu, P. Mika, and N. Torzec. Entity Recommendations in Web Search. In *Proceedings of the 12th International Semantic Web Conference - Part II, ISWC '13*, pages 33–48, New York, USA, 2013. Springer.
- [3] A. Blumauer. From Taxonomies over Ontologies to Knowledge Graphs, July 2014. <https://blog.semantic-web.at/2014/07/15/from-taxonomies-over-ontologies-to-knowledge-graphs> [August, 2016].
- [4] J. Davies, R. Studer, and P. Warren. *Semantic Web Technologies: Trends and Research in Ontology-based Systems*. John Wiley & Sons, 2006.
- [5] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmman, S. Sun, and W. Zhang. Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14*, pages 601–610, New York, USA, 2014. ACM.
- [6] J. Euzenat and P. Shvaiko. *Ontology Matching*. Springer, Secaucus, NJ, USA, 2007.
- [7] M. Färber, B. Ell, C. Menne, A. Rettinger, and F. Bartscherer. Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web Journal*, 2016. <http://www.semantic-web-journal.net/content/linked-data-quality-dbpedia-freebase-opencyc-wikidata-and-yago> [August, 2016] (revised version, under review).
- [8] M. Färber and A. Rettinger. A Statistical Comparison of Current Knowledge Bases. In *Joint Proceedings of the Posters and Demos Track of 11th International Conference on Semantic Systems - SEMANTiCS2015 and 1st Workshop on Data Science: Methods, Technology and Applications (DSci15)*, pages 18–21. CEUR Workshop Proceedings, 2015.
- [9] C. Feilmayr and W. Wöß. An Analysis of Ontologies and their Success Factors for Application to Business. *Data & Knowledge Engineering*, 101:1–23, 2016.
- [10] P. James. Knowledge Graphs. In *Linguistic Instruments in Knowledge Engineering*, pages 97–117. Elsevier Science Publishers B.V., 1992.
- [11] S. Krause, L. Hennig, A. Moro, D. Weißenborn, F. Xu, H. Uszkoreit, and R. Navigli. Sar-graphs: A Language Resource Connecting Linguistic Knowledge with Semantic Relations from Knowledge Graphs. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, Special Issue on Knowledge Graphs(C):112–131, Mar. 2016.
- [12] M. Kroetsch and G. Weikum. Journal of Web Semantics: Special Issue on Knowledge Graphs. <http://www.websemanticsjournal.org/index.php/ps/announcement/view/19> [August, 2016].
- [13] Y. Ma, P. A. Crook, R. Sarikaya, and E. Fosler-Lussier. Knowledge Graph Inference for Spoken Dialog Systems. In *Proceedings of 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015*. IEEE - Institute of Electrical and Electronics Engineers, April 2015.
- [14] P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. A. Knoblock, D. Vrandečić, P. T. Groth, N. F. Noy, K. Janowicz, and C. A. Goble, editors. *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, volume 8796 of *Lecture Notes in Computer Science*. Springer, 2014.
- [15] S. Nurdiati and C. Hoede. 25 Years Development of Knowledge Graph Theory: The Results and the Challenge, September 2008.
- [16] H. Paulheim. Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. *Semantic Web Journal*, (Preprint):1–20, 2016.
- [17] J. Pujara, H. Miao, L. Getoor, and W. Cohen. Knowledge Graph Identification. In *Proceedings of the 12th International Semantic Web Conference - Part I, ISWC '13*, pages 542–557, New York, USA, 2013. Springer.
- [18] A. Singhal. Introducing the Knowledge Graph: Things, not Strings, May 2012. <https://googleblog.blogspot.co.at/2012/05/introducing-knowledge-graph-things-not.html> [August, 2016].
- [19] T. Steiner, R. Verborgh, R. Troncy, J. Gabarró Vallés, and R. Van de Walle. Adding Realtime Coverage to the Google Knowledge Graph. In *Poster and Demo Proceedings of the 11th International Semantic Web Conference*, Nov. 2012.
- [20] F. M. Suchanek and G. Weikum. Knowledge Bases in the Age of Big Data Analytics. *Proceedings of the VLDB Endowment*, 7(13):1713–1714, Aug. 2014.
- [21] A. Tonon, M. Catasta, R. Prokofyev, G. Demartini, K. Aberer, and P. Cudré-Mauroux. Contextualized Ranking of Entity Types Based on Knowledge Graphs. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, Special Issue on Knowledge Graphs(C):170–183, Mar. 2016.