Ontology Summit 2020 Communiqué: Knowledge Graphs

Ken Baclawski, Northeastern University, Boston, MA USA
Michael Bennett, Hypercube Limited, London, UK
Gary Berg-Cross, ESIP Semantic harmonization Co-Lead
Todd Schneider, Engineering Semantics, Fairfax, VA USA
Ravi Sharma, Senior Enterprise Architect, Elk Grove, CA USA
Janet Singer, INCOSE, Scotts Valley, CA USA
Ram D. Sriram, National Institute of Standards Technology, Gaithersburg, MD USA

Abstract: An increasing amount of data is now available from public and private sources. Furthermore, the types, formats, and number of sources of data are also increasing. Techniques for extracting, storing, processing, and analyzing such data have been developed in the last few years for managing this bewildering variety based on a structure called a knowledge graph. Industry has devoted a great deal of effort to the development of knowledge graphs, and knowledge graphs are now critical to the functions of intelligent virtual assistants such as Siri, Alexa, and Google Assistant. The goal of the Ontology Summit 2020 was to understand not only what knowledge graphs are but also where they originated, why they are so popular, the current issues, and their future prospects. The summit sessions examined many examples of knowledge graphs and surveyed the relevant standards that exist and are in development for knowledge graphs. The purpose of this Communiqué is to summarize our understanding from the Summit in order to foster research and development of knowledge graphs.

Keywords: knowledge graph, knowledge graph architecture, ontology, semantics, semantic networks

1. Introduction

While there is a long history of the use of knowledge graphs (KGs) across various domains, they have proven in the last few years to be an especially important tool for semantic technology and research areas. As structured representations of semantic knowledge that are stored in a graph, KGs are lightweight versions of semantic networks that potentially scale to massive data repositories such as the entire World Wide Web ("Semantic Network", 2020). Industry has devoted a great deal of effort to the development of knowledge graphs, and they are now critical to the functions of intelligent virtual assistants such as Siri, Alexa, and Google Assistant. Some of the research communities where KGs are relevant include applied ontology, big data, linked data, Open Knowledge Network (OKN), artificial intelligence (AI), and deep learning.

The Ontology Summit 2020 examined KGs from several perspectives during a series of virtual sessions held from September 2019 to June 2020. This Communiqué synthesizes and summarizes the findings of this series. The Ontology Summit 2020 was organized by question words whose answers are considered basic for information gathering, problem solving, or establishing a context. These questions include the traditional Five Ws to which we added "How", "Whence", and "Whither", as shown in Figure 1. Accordingly, this Communiqué is organized based on these question words. In order to promote a consistent terminology for the notion of a KG, we begin in Section 2 by proposing a practical answer to "What" with a definition of a KG based on definitions published in the literature as well as on invited speakers and discussions during the summit. Section 3 gives some suggestions for "Why" KGs have recently begun to be popular, as well as from "Whence" KGs arose. Section 4 is devoted to addressing "How", "Who", "Where", and "When" by examining examples of techniques and tools used by the many activities of KG systems. Sections 5, 6, and 7 are concerned with "Whither". Section 5 lists standards and standardization efforts relevant to KGs, and Section 6 lists some of the problems and challenges of KGs that were identified by the Ontology Summit. Section 7 speculates about the future prospects of KGs. The Communiqué ends with a conclusion and acknowledgments.

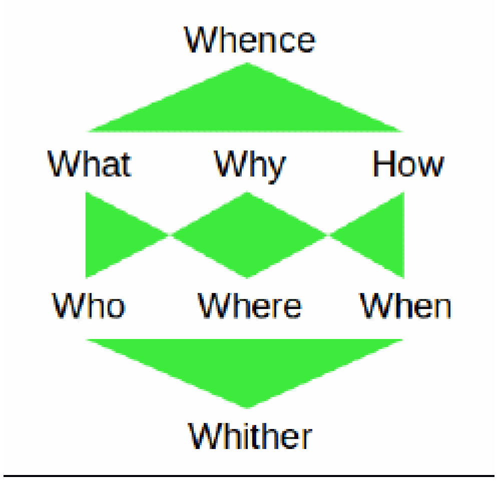

Figure 1: The Context Questions

The Ontology Summit 2020 covered a great deal of material. This Communiqué only outlines the main points of the 32 sessions that occurred over 9 months. Consequently, much of what took place is not covered in this article. It is our intention to present the findings of the summit more completely and in more detail in a series of articles to be published separately.

2. What is a Knowledge Graph?

We begin by addressing the question of what a knowledge graph is. Unfortunately, there are a great many academic papers as well as websites and companies that have proposed many different definitions. To synthesize a coherent definition that helps frame the discussion about KGs, the definitions in references (Krötzsch and Thost, 2016; Paulheim, 2017; Blumauer, 2014; Färber, Ell, Menne, Rettinger, and Bartscherer, 2018; Pujara, Miao, Getoor, and Cohen, 2013; Rohrseitz, 2019; Aijal, 2019; Bergman, 2019; Aasman, 2019) were reviewed. They had the following common features:

1. A KG represents interrelationships. All of the definitions specify this feature but do so in different ways.

2. A KG uses techniques to extract knowledge from one or more sources. The kinds of sources differ from one definition to another.

3. The organization is a graph, although the precise meaning of "graph" varies from one definition to another.

4. While a KG must have a schema, not all KG definitions mention it. Those that do mention it specify that the schema defines classes and relations.

5. The KG supports various graph-computing, search, and query interfaces. The supported operations and performance will vary, and the performance will depend on how trade-offs among scalability, performance, and maintainability are handled as well as on other technical issues.

From these features it is apparent that a KG is not simply another way to represent facts. It involves a software architecture that includes active capabilities for extracting and processing the facts. Jans Aasman (2019) characterized the operations of a KG as follows:

- Generation:
  - Collection: Ingestion, web extraction, catalog extraction, ontology, ...
  - Processing: Schema mapping, entity resolution, cleaning, ...
- Storage
- Applications: Querying, graph mining, recommendation, search, question answering, ...
- Statistical and machine learning techniques are used for all of the above

Another example of a definition was given by Nicola Rohrseitz as follows:

> "A Knowledge Graph is a set of datapoints linked by relations that describe a domain, for instance a business, an organization, or a field of study. ... Knowledge Graphs are secondary or derivate datasets: They are obtained by analyzing and filtering the original data. ... Knowledge Graphs are also sometimes called semantic networks. Semantic emphasizes the fact that the meaning is encoded together with the corresponding data. This is done through the taxonomies and ontologies ..." (Rohrseitz, 2019)

Rohrseitz went on to describe how KGs are constructed and used. This description is similar to the characterization of KGs given by Jans Aasman above.

The fact that it is the operations of a KG that are its primary distinguishing feature may be the reason for the confusion about what a KG is. It has also led many to characterize KGs as being "nothing new" and as simply another buzzword. Unfortunately, the term "knowledge graph" is partly to blame for this. It tends to suggest that a KG is no more than a special kind of graph or network. Accordingly, it might be helpful to employ a term that is less confusing such as "knowledge graph architecture" (KGA), which is defined below. Figure 2 is one example of a KGA.

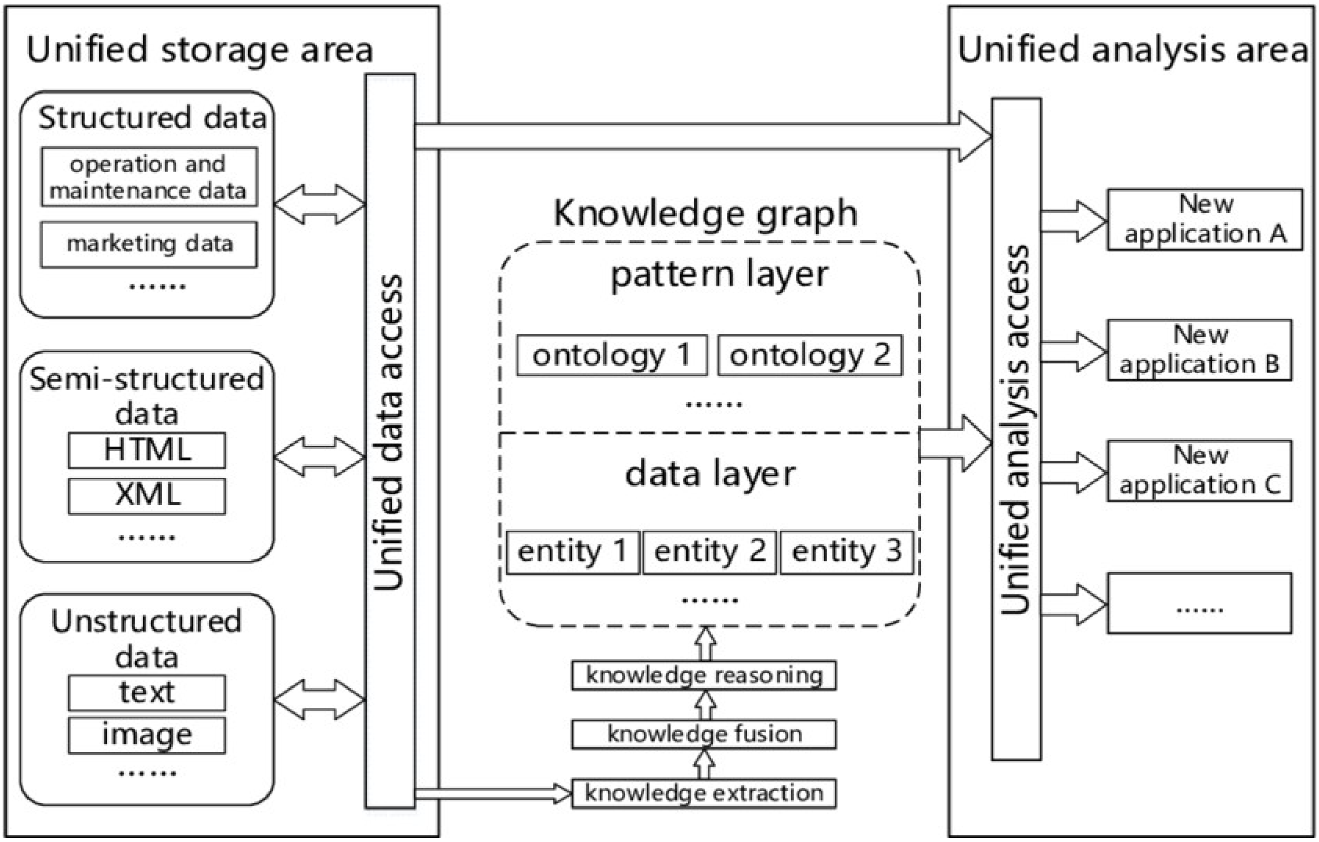

Figure 2: A Knowledge Graph Architecture from (Yuan, Zhang, Dai, Peng and Zhao, 2018)
HTML is the Hyper Text Markup Language and XML is the Extensible Markup Language

The architecture for a KG system is analogous to the architecture for a data warehouse (DW). DW technologies have been used to integrate and harmonize data so that analysts and users can reliably extract meaning from their large enterprise datasets. But while well-established, the DW approach involves significant up-front and ongoing costs, as well as serious risks. Further, due to data complexity, DWs don't address significant areas of enterprise data. However, to be fair, the KG approach also can have significant costs and complexity. Data warehousing leverages older technologies that lack the flexibility of KGs, making them too slow to meet the ever-changing demands of Big Data. KG systems offer a more modern, flexible and dynamic approach to data sharing and integration; and, as discussed in Section 4, many different methods and technologies are employed by KG systems.

These findings lead to the following proposal for a definition of a KG and KGA:

> A KG is a representation of a set of statements in the form of a node- and edge-labeled directed multigraph allowing multiple, heterogeneous edges for the same nodes. A collection of definitional statements specifying the meaning of the knowledge graph's labels is called its schema.
>
> A KGA provides a combination of scalable technologies, specifications, and data cultures for representing densely interconnected statements derived from structured or unstructured sources across domains in a reasonable way that is both human- and machine-readable.
>
> A KGA together with a collection of KGs is a KG system (KGS).

A "node- and edge-labeled directed multigraph" is an 8-tuple $(V, E, s, t, \Sigma_V, \Sigma_E, \ell_V, \ell_E)$ such that

1. V is a set of nodes and E is a set of edges.
2. $s\colon E \longrightarrow V$ and $t\colon E \longrightarrow V$ are functions that specify the source and target nodes of the edges.
3. Σ_V is a set of node labels, and Σ_E is a set of edge labels.
4. $\ell_V\colon V \longrightarrow \Sigma_V$ and $\ell_E\colon E \longrightarrow \Sigma_E$ are functions that specify the labels of the nodes and edges.

As KGs and KGAs are engineering artifacts, they have associated processes for development, testing, validation, management, and the overall lifecycle. By analogy with DataOps (Liebmann, 2020), the set of practices that combines all of these processes with the KGS might be called the KnowOps. Despite the name "knowledge graph", there is no requirement that the statements be implemented as a graph. A collection of KG systems could themselves be the sources for an overarching KG that fuses the source KGs.

While ontologies are not specified in the definition of a KG, they can play an important role even though this role varies. Some KGs incorporate an ontology as part of the structure, in which case the notions of KGs and ontologies are essentially equivalent. In other cases, the KG and ontology are decoupled, and it is possible for one KG to have more than one associated ontology so that an ontology plays the role of a relational database view.

3. Why use Knowledge Graphs?

We now examine the question of why KGs and KG systems have become popular. This is part of the broader question of why one bothers with information at all, a question provocatively asked by Matthew West (West, 2020). Within that broader context, Jans Aasman suggested several reasons why KGs have recently become so popular (Aasman, 2019).

In business, information is used to support decisions. If information required for a decision is missing or inaccurate, the risk of a mistake increases. So, to support a decision, information needs to be fit for purpose, which means information management is a quality management process where information is the product.

But how does one know what the information requirements are? It turns out that asking people for their requirements gives unreliable results. A better approach is to document the processes to the level where key decisions are explainable. It is then possible to document the information requirements for those decisions.

Information has a lot of properties, but only some of them are critical for its use in supporting decisions. One of the hardest properties to achieve is consistency. If data is consistent, then when it arrives from different sources it can just be brought together and used immediately. Consistent data uses the same data model and reference data (or, if you prefer, knowledge graphs of the same ontology). However, if the sources are not consistent, either individually or with each other, then one must not only extract information from sources but also resolve the inconsistencies. Consequently, it is necessary to develop a set of tools for this purpose. In other words, there is a need for a software architecture for the information.

Given that one needs a system for capturing knowledge, a natural question is what it is that made KGs so popular. While one can only speculate about the reasons, the following are plausible explanations:

- Graph databases are now accepted as the best technology to store complex semantic data.

- People are no longer afraid of taxonomies, although ontologies are still intimidating.
- Entity extraction and Natural Language Processing (NLP) are almost a commodity now with spaCy, Bidirectional Encoder Representations from Transformers (BERT), IBM Natural Language Understanding, and many other tools.
- Machine learning and advanced analytics are now available in the cloud. (Aasman, 2019)

Note that a capability for reasoning/inference is not in this rationale. Indeed, there are successful KG systems that either have a minimal schema or do not have significant emphasis on the schema. That said, there is general agreement about the usefulness of ontologies for KG systems.

4. Knowledge Graph Techniques and Tools

In this section we provide a sample of the kinds of techniques and tools being used in and being developed for KGs. Section 4.1 describes various forms of reasoning and mathematical techniques from probability theory and category theory for KGs. Then, in Section 4.2, we describe the Open Knowledge Network, a U.S. National Science Foundation program for KGs that is supporting some of the projects in the subsequent subsections. One important challenge for KGs is spatial and temporal reasoning, and Section 4.3 presents two projects addressing this. The rest of this section is devoted to projects in some of the many domains for which KG techniques have been applied. Section 4.4 is concerned with extracting KGs from scientific publications, Section 4.5 is concerned with KGs in product design and manufacturing, Section 4.6 describes two applications of KGs to government problems, and Section 4.7 proposes to use KGs for a new kind of dynamically interactive textbook. For more details about each project, please see the link to the associated slide or video presentation given by the cited reference.

4.1 Techniques

Despite the varying definitions of the notion of a KG, there is a common goal: to use these KGs to gain important insights and make data-driven discoveries. Anirudh Prabhu defines “insights” as important patterns, trends, and concordant information obtained from the knowledge graphs, especially in cases where such features are not obvious from simple data exploration tasks (Prabhu, 2020). Using reasoners to gain insights and make inferences about the data is a method commonly known and utilized in the Semantic Web community. But by utilizing methods (both visual and analytical) known in network science, one can identify previously unseen patterns and trends and use these insights to generate or validate hypotheses and aid in scientific discoveries. Global metrics are used to gain insights about the entire network structure and compare two or more networks with each other. Local metrics are used to inspect individual network structure and find important trends within that network. Community detection algorithms are used to mathematically identify groups of nodes within a network, usually based on how these nodes are connected to each other. Lastly, Prabhu examined (both visually and mathematically) the evolution of a network based on the change in a specific data feature (e.g., time, pressure, or temperature) to identify how the addition or removal of a node (or set of nodes) affects the overall network structure.

Another approach to gain insights is to use probabilistic KGs as presented by (Srihari, 2020). These KGs incorporate statistical models for relational data. Triples are assumed to be incomplete and noisy. There are two main types of models: latent feature models and Markov random fields (MRFs). Latent feature models can be trained using deep learning. MRFs can be derived from Markov Logic Representations of facts in a database.

Yet another technique for gaining insights is to use the mathematical theory of categories and functors. In "Composing Knowledge Graphs, inside and out", Spencer Breiner explained how some of the limitations of graph-based knowledge representations can be addressed formally by using foundational methods from category theory (Breiner, 2020). While category theory is regarded as very abstract even among mathematicians, categories are in fact closely related to KGs. A category consists of a collection of objects and arrows (directed links) between them, which is exactly what one means by a directed graph. This approach can be applied to practical issues. To illustrate its use for a practical issue, the problem of open-shop scheduling in operations research was presented using category theory.

4.2 The Open Knowledge Network Program

OKN is a program of the U.S. National Science Foundation whose goals include the development of the following:

- An advanced science data infrastructure that is interoperable and has an open architecture, making it easier to access and link heterogeneous data products
- An open semantic information infrastructure to discover new knowledge from multiple disparate knowledge sources
- A nonproprietary shared knowledge infrastructure, with a particular focus on publicly available data, e.g., U.S. government, scientific data, and other similar public datasets (Baru, 2020)

The OKN benefits multiple applications domains, including science and engineering research. More succinctly, it is “A Siri for Science." The common themes for sponsored OKN projects include:

- Integrating heterogeneous types of data
- Accommodating dynamic information
- Supporting access by and contributions to the KG by heterogeneous communities of users
- Incorporating new information into the knowledge graphs using machine learning and crowdsourcing approaches

Some of the projects described below are sponsored by the OKN program.

4.3 Temporal and Spatial Projects

One challenge facing KGs is the problem of representing time and space. Even very powerful AI systems can falter in dealing with time. If you ask Google “How old is Joe Biden?” or “How old is Mitch McConnell?”, you get the correct answers; but if you ask “Who is older, Joe Biden or Mitch McConnell?”, all you get are links to articles that mention both politicians. The problem is that while KGs typically do include temporal features of entities, they are treated as little more than textual strings with no other semantics. Furthermore, many features and relations that are in fact time-dependent, such as the spatial extent of countries, are treated as timeless. This situation is surprising since temporal reasoning is highly developed in AI and database management. Some aspects of temporal and spatial reasoning were covered in the Ontology Summit 2018 on Context and Ontologies (Baclawski et al, 2018). Additionally, standards bodies have been developing temporal and spatial representation and reasoning standards as discussed in Section 5. Unfortunately, within KGs time is an afterthought if it is included at all. Spatial reasoning has similar challenges, although the need for spatial reasoning is less common than for temporal reasoning. The KG research community should explore all aspects of time and space, from the abstract to the concrete, from general purpose reasoning to highly specific applications. In the long run, the benefits to AI systems of effective, flexible temporal and spatial reasoning will be large (Davis, 2020). The following two projects are attempting to respond to this challenge.

The KnowWhereGraph developed by Krzysztof Janowicz is a project that takes a geographic information system (GIS) to the next level, by providing open graph-based linking and semantic enrichment technologies far beyond pre-defined data themes and silos (Janowicz, 2020). The ultimate goal is to understand how to engineer meaningful features (independent variables) via a KG-based GIS for downstream models such as supply chain forecasting or soil health mapping by including spatial-temporal semantics.

Sean Gordon is part of a team that is prototyping an OKN for spatial decision support (Gordon, 2020). Based on existing work by members of the team, four case study sub-teams were created that are working on needs analysis for multi-stakeholder organizations focused on three core environmental themes (water quality, wildland fire, biodiversity) in different regions of the western U.S.; and one case study sub-team was created that is working on a professional body of knowledge for geographic information science and technology. Each of the four case study sub-teams used interviews and/or workshops to have collaborators identify need-to-know-concerns and questions. This approach helped prioritize a) schema for a KG that will support decision making for each theme; b) spatial decision support resources to add to the KG; and c) particular use cases.

4.4 Scientific Publishing

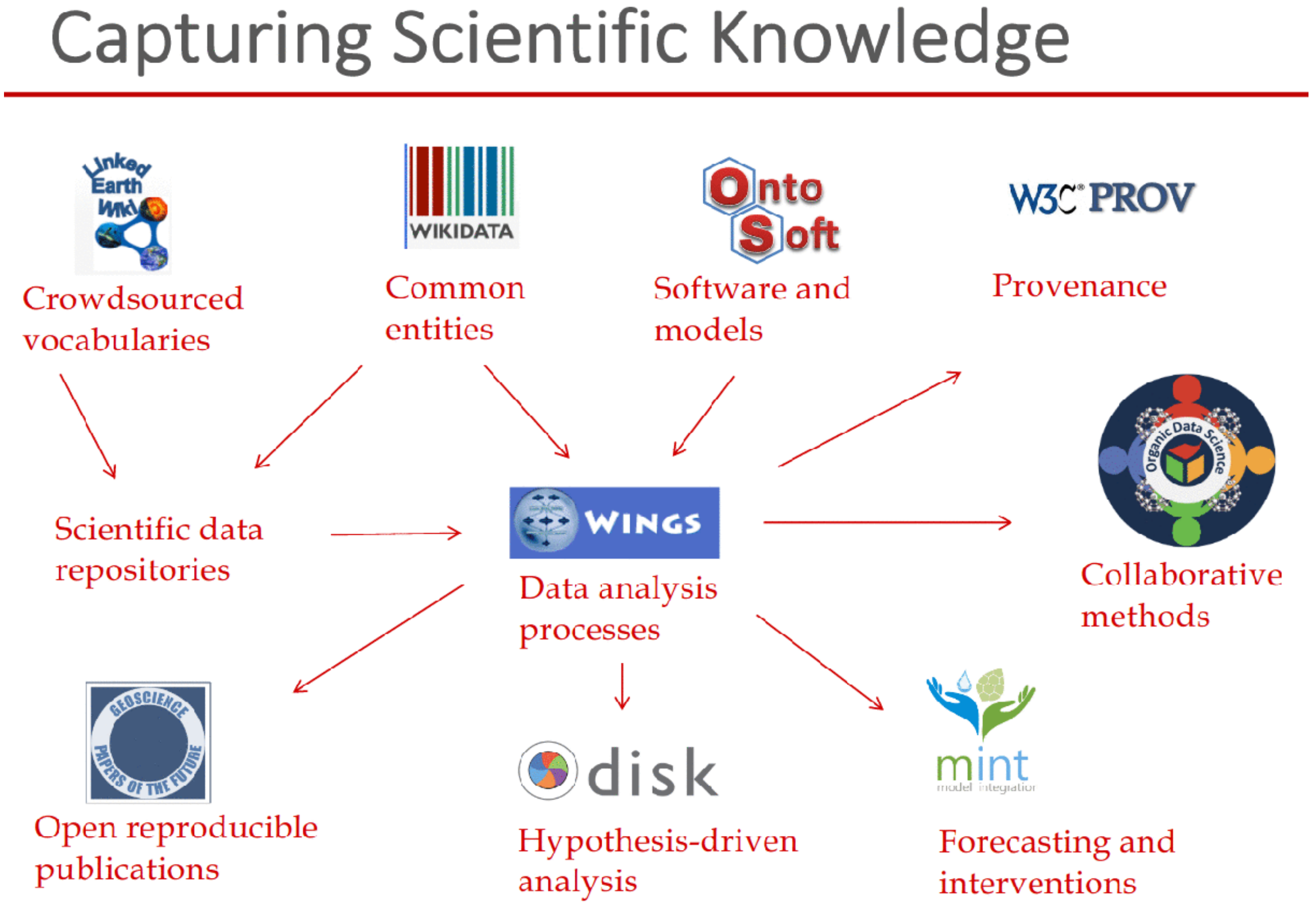
Capturing Scientific Knowledge
Linked Earth Wiki
Crowdsourced vocabularies
WIKIDATA
Common entities
Onto Soft
Software and models
W3C PROV
Provenance
Scientific data repositories
WINGS
Data analysis processes
Organic Data Science
Collaborative methods
GEOSCIENCE PAPERS OF THE FUTURE
Open reproducible publications
disk
Hypothesis-driven analysis
mint model integration
Forecasting and interventions

Figure 3: Links among entities of scientific knowledge from (Gil, 2020)

The main output of Science is publications. There are around 30 000 journals, and about two million papers are published every year. Efforts to extract the knowledge in the scientific record predate the

World Wide Web (Baclawski, Futrelle, Fridman, and Pescitelli, 1993; Baclawski et al, 1993). Yolanda Gil describes seven ontologies that provide essential capabilities, but much work remains to be done to capture more comprehensively the scientific record. Are we far from a day when each scientific article will be properly linked to hypotheses, models, software, provenance, workflows, and other key scientific entities on the Web as shown in Figure 3? Will AI research tools then be able to access this information to generate new results? Will AI systems ultimately be capable of autonomously writing scientific papers in the future? (Gil, 2020)

4.5 Manufacturing

The mission of the Manufacturing Open Knowledge Graph (MOKN) project is to structure the world's public information on product design and manufacturing (Starly, 2020). MOKN's broader impact is to make information available regarding sourcing critical part components, instantaneous gathering of specific manufacturing capabilities, location of those services, and availability of resources. The global pandemic crisis serves as a contextual example as to the value of this knowledge especially for alternate sourcing and prequalification of vendors – with implications for public health and national security. Accessibility also empowers rural and suburban communities dependent on manufacturing services.

4.6 Government

Matthew West is involved with an ambitious attempt in the UK to develop a Digital Twin of the entire national infrastructure. The aim is to establish a distributed Digital Twin of consistent data so that authorized users can construct queries across the Digital Twins in order to answer questions like "Which Tower Blocks have the same type of cladding as Grenfell Tower?" The Information Management Landscape sets out the information needed to support the critical properties of data and the information quality management process. Part of this infrastructure is an integration architecture that allows the distributed National Digital Twin to be virtualized so users can see it as a single database with access only to the data they are authorized to see (West, 2020). In effect, this is a KG system for which the underlying source data is extracted from a large collection of KG systems, each of which is devoted to a single city or small region.

The Rich Context project described by Paco Nathan is the KGA of the Administrative Data Research Facility (ADRF) platform is currently used by 50 federal, state, and local agencies in the U.S. to identify people with specific expertise (Nathan, 2020). ADRF was cited as the first example of Secure Access to Confidential Data in the final report of the Commission on Evidence-Based Policymaking.

4.7 Education

College students today face the challenge of mastering concepts in new subject areas and relating those concepts across multiple disciplines, yet their textbooks have the "one size fits all" nature. In "Textbook Open Knowledge Network", Vinay K. Chaudhri presented Intelligent Textbooks (ITB) using AI and KGs to solve these problems. Students can dynamically interact with the textbook content, increasing their ability to understand concepts, increasing engagement, and thereby, improving academic performance (Chaudhri, 2020).

5. Standards

We now discuss some standards that are relevant for KGs. What makes standards especially useful for KGs is that there are significant distinctions among the many KGs that have been developed. Standards can help such disparate KGs to interoperate. Standards also serve the purpose of KG development. For example, one can develop a standard to represent objects and relationships for a manufacturing KG, which can be used all over the world to develop KGs in a particular domain. These KGs can then be more easily integrated at a later stage. KG systems differ not only in the sources for their knowledge (e.g., the Web, sensor data in some domains, commercial transaction data, etc.) but also in the operations for generating, processing, and utilizing the results. For example, does the KG system support reasoning? If it does, then what kind of reasoning? Is an entire KG accessible for reasoning? When reasoning or inferencing is used, there is an expectation that the result of such action will produce results that are consistent with expected interpretation(s). Such interpretation(s) are based on distinctions among the entities involved in the inference and are expressed via the (usually, natural language) symbols (also known as labels) used in the representation.

A KG is created to meet certain needs and uses in some context, though the context may not be sufficiently or explicitly recognized (or represented). Consequently, a KG will necessarily have limitations of coverage (i.e., scope) and completeness (level of detail), which will hinder interoperability. There are several ways to deal with this problem. One option is to employ an ontological analysis when creating the KG. Related to that is the use of a well-developed ontology as the schema based on such an analysis. Another option is to use applicable standards (e.g., engineering, terminology, reasoning, etc.) which is the subject of this section.

In a paper on the role of standards in innovation, Allen and Sriram state "Standards are documented agreements containing technical guidelines to ensure that materials, products, processes, representations, and services are fit for their purpose" (Allen and Sriram, 2000). They then discuss how standards introduced at the right time will lead to greater innovation. For example, the standardized musical notation has spurred hundreds of years of creative music compositions.

Lisa Carnahan further elaborated on standards and the process of creating standards in her Ontology Summit talk "The IT Standard Process" (Carnahan, 2020). In the U.S., standards are developed by standards developing organizations (SDO). An SDO is any organization that develops and approves documented standards using various methods to establish consensus among its participants. There are hundreds of SDOs. Such organizations may be: accredited (e.g., the International Committee for Information Technology Standards is accredited by the American National Standards Institute); international treaty-based (e.g., International Telecommunication Union-Telecommunication, International Civil Aviation Organization); international private-sector based (e.g., International Organization for Standardization (ISO), Institution of Engineering and Technology (IEC) or the Institute of Electrical and Electronics Engineers (IEEE)); an international consortium (e.g., Object Management Group (OMG), Organization for the Advancement of Structured Information Standards (OASIS), Internet Engineering Task Force (IETF), or World Wide Web Consortium (W3C)); or a government agency (e.g., Department of Defense, Department of Homeland Security, National Institute of Standards & Technology (NIST)).

One of the SDOs is the ISO, the world's largest developer of voluntary international standards. Barry Smith spoke at the Ontology Summit about his experiences with improving interoperability of KGs, emphasizing ISO/IEC 21838 (Smith, 2020). This standard is titled "Information technology — Top-level ontologies" and includes as one of its parts, the Basic Formal Ontology (BFO). Ontologies have been enormously successful in the biomedical field for some 20 years, where the Gene Ontology (GO), the first version of which was created in 1998, was referred to from the very beginning as a 'directed acyclic graph' representing knowledge about genes and gene products. The foundational ontology of GO is the BFO. With the growth in impact of the data from the human and other model organism genome projects, the data-annotation needs of the biomedical informatics world expanded tremendously, and this led to the creation of new ontologies, for example for proteins, cell types, diseases, and others. This expansion of ontology development continues to this day with the new COVID-19 ontology. The influence of BFO in non-medical domains is indicated also by the reception of the ISO/IEC 21838 standard in areas such as digital manufacturing, particularly through the creation of the Industrial Ontologies Foundry (IOF). Under the auspices of this entity, work is on-going to relate BFO to current developments on the Standard for the Exchange of Product model data (ISO 10303) and the manufacturing technology standard (MTConnect) for factory device data ("Industrial Ontology Foundry", 2020).

The W3C is the main SDO for the World Wide Web. The W3C standards that are most closely related to KGs are the Resource Description Framework (RDF), RDF Schema (RDFS) and the Web Ontology Language (OWL). RDF is a graph-based modeling language, while RDFS and OWL are ontology languages layered on RDF.

Another SDO is the OMG that is best known for the Unified Modeling Language (UML), the Meta-Object Facility (MOF), and the Model Driven Architecture (MDA). Elisa Kendall provided an update on OMG standards and activities that are relevant to ontologies and KGs (Kendall, 2020). The Ontology Platform Special Interest Group (OPSIG) has been an active, contributing working group for 15+ years. So far, the following Platform standards have been published: Ontology Definition Metamodel (ODM); Distributed Ontology, Model and Specification Language (DOL); Languages, Countries and Codes (LCC); and the MOF to RDF Mapping – MOF2RDF. The OMG has also published several domain-specific ontologies, including the Financial Industry Business Ontology (FIBO), Financial Instrument Global Identifier (FIGI), and the Information Exchange Packaging Policy Vocabulary (IEPPV). A Robotic Service Ontology is now being prepared jointly with the IEEE Robotics community.

Other standards that are related to KGs include Case Management Model and Notation (CMMN), Decision Model and Notation (DMN), Date Time Vocabulary (DTV), Production Rule Representation (PRR), and Semantics of Business Vocabularies and Rules (SBVR). Still others are in preparation. However, there will not be one gold standard for KGs. Several standards will emerge and will need to be judiciously mixed.

Common Logic (CL) is the ISO/IEC standard for First-Order Logic (ISO/IEC 24707:2007). In "Knowledge Graphs and Logic", John Sowa gave an overview of CL and related standards for logic (Sowa, 2020). The CL standard includes specifications for three dialects: the Common Logic Interchange Format (CLIF), the Conceptual Graph Interchange Format (CGIF), and an XML-based notation for Common Logic (XCL). The CLIP dialect combines the best features of two dialects, CLIF and CGIF. The primary design goals for CLIP are the following:

- Immediately readable by anyone who knows predicate calculus
- As readable as Turtle for the RDF and OWL subsets
- As readable as any notation for if-then rules
- Serve as a linearization for a wide range of graph logics, including conceptual graphs, existential graphs, KGs, RDF, OWL, and UML diagrams
- Query option: Select (list of names) where (any CLIP sentence)
- Support mappings between logics and natural languages

The DOL standard mentioned above is an OMG standard for integration and interoperation among distributed ontologies, models, and specifications (OMS). DOL is formally defined by logic and mathematics. In other words, DOL can integrate heterogeneous OMS by relating the logics that specify them.

The financial services industry is an exceptionally large, mature, and data-intensive industry that has an impact on virtually everybody. Michael Bennett presented an overview of KGs in the financial sector (Bennett, 2020). While most of the historical standards in the financial services industry deal with messaging requirements or data formats, there are also industry standards for formal semantics. FIBO was conceived to provide a common language across these messaging standards, while a more recent initiative from the ISO Technical Committee dealing with financial services aims to supplement the existing ISO 20022 XML messaging standard with formal semantics. FIBO arose out of a need to harmonize terms across the industry as a common language for reuse of data in reporting, risk management, and compliance. This need arose out of a realization that, while it was hard to reach agreement on common terms, the concepts themselves were well understood.

While the Financial Industry is a specific domain, it provides important lessons that are relevant to ontologies and KGs in general. For example, one distinction is whether to provide a deep hierarchy of foundationally primitive terms based around a Top-Level Ontology (TLO) or not. These are typically not needed for OWL applications and have been removed from the OMG FIBO standard. Another distinction is whether the ontology represents real-world 'truth-makers' (assertions that give rise to the meaning of a class of things) or data about things. For example, to ***be*** a bank is to hold certain legal capacities and capabilities, whereas to ***know something is*** a bank is to interrogate the available data for some suitable 'data signature' that such capacities exist, in this case in the form of a banking license. Ontologies therefore may be foundational for use as a point of reference or may be application-focused; and they may be predicated on subject matter or on data about that subject matter. These distinctions may be dismissed by developers as unimportant, but if one does not address them the result is that interoperability can be severely inhibited.

In "Standards and Ontologies" Michael Grüninger discussed the advantages and disadvantages of standardization of ontologies (Gruninger, 2020). The problem with de facto standards is that ontologies will be adopted simply because they are popular and widely used even if they were not properly developed with sufficient evaluation and analysis. The risk with this approach is ontologies could be used that contain ontological errors, unintended models, and omitted models, or they could incorporate implicit ontological commitments that prevent reuse. The standards we need are, therefore, the ones that enable the evaluation and comparison of ontologies. First of all are standards for ontology representation languages with formal semantics such as Common Logic (ISO 24707) and OWL. Second are standards for the specification for mappings between ontologies and between logics, a prime example being DOL from OMG. Finally, there are standardized axiomatizations of ontologies, in particular ISO 18629 (Process Specification Language) and ISO 21838 (Top-Level Ontologies).

Much work remains to be done in the standards arena. Recently, the International Association of Ontology and its Applications (IAOA) established the Industry and Standards Technical Committee (ISTC). This committee has two core purposes:

1. To foster the use of applied ontology in standardization initiatives,
2. To facilitate the interactions across people in industry and in applied ontology research.

Activities within the ISTC include the dissemination of information about initiatives with the aim to gather experts interested in the development of ontologically sound standards. The ISTC also organizes virtual and physical meetings and events to discuss how to understand and apply ontological approaches and methodologies, both in general and for KG systems in particular.

6. Challenges

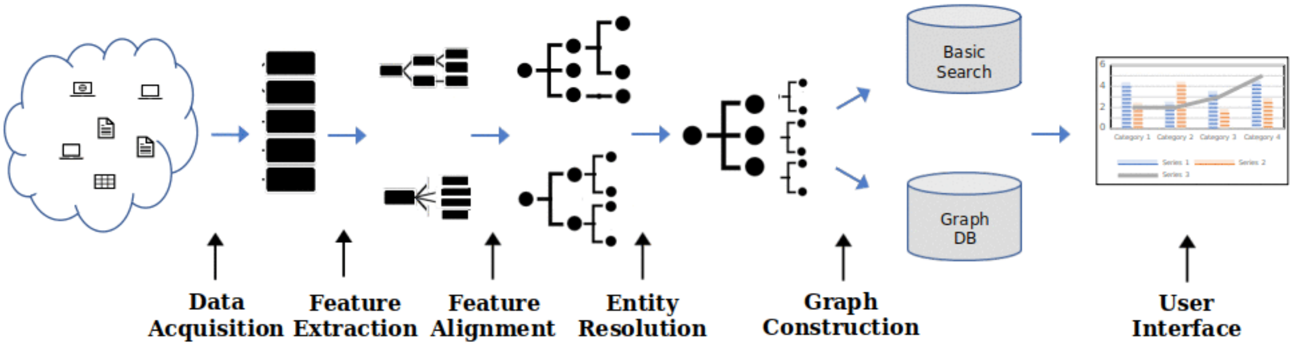

Figure 4: A Pipeline for Building a KG

Methods of building a KG take us from raw, messy, and disconnected data/information that is hard to query, analyze, and visualize to a more refined, organized, cleaned, and linked product that is easier to visualize, query, and analyze. Challenges exist at every step in this process including recursions as part of a life cycle. In this section we briefly list these challenges in Table 1. The first column in Table 1 is an operation of a KGA. These operations are labeled "KG Step" because they are usually the steps in a pipeline of operations such as that shown in Figure 4. For each KG step there may be many problems and issues. We list the most significant of these in the second column of Table 1. The next column describes the context of the problem or issue. The last column cites some references. More details about these challenges will be published in a separate article.

Table 1: Knowledge Graph Challenges

KG Step	Problem/Issue	Context	Notes	Reference
Scoping	Identifying best available sources in the vast space of possibilities	Understanding usage and knowledge requirements	Determining which candidate facts should be included into a KG.	(Pujara, Miao, Getoor, and Cohen, 2013)
Data Acquisition and populations	Data volume, variety, speed, and validity. There may be too little structured data to seed a graph	The space of needed data may be unknown. Domain metadata and interdomain metadata evolve over time.	We will not know all the types and relationships needed to model. Need to explore entity neighborhoods and compare neighbor entities and values.	(Dong, 2020)
Feature Extraction	Need mature processes to find types & create vector features usable by	Besides technical challenges, is there a need to validate extractions to	Active learning, weak learning, distant supervision along with semi-supervised learning, transfer	(Joshi, 2019; Dong et al, 2020; Wang, Xu, Li, Dong, Gao, 2020)

	machine learning models. Limited training labels for large-scale, rich data. Often hidden patterns, such as in headings, carry key relations, attributes, dates, etc.	understand if extracted features are aligned to human judgment?	learning, and meta-learning are all techniques to address limited training data. Mentalistic terminology of features may be misleading to those outside of computer science. Important info is also in images, making features harder to extract.	
Feature Alignment	Heterogeneous data and large data spaces challenge alignment across many records.	Are two features the same? Do “born” and “date of birth” mean the same?	Explore redundancy of data. Are values of the same attribute in the same embedding space?	(Pham, Alse, Knoblock, and Szekely, 2016; Taheriyan, Knoblock, Szekely, and Ambite, 2016)
Entity Resolution	Noisy data problems and scale. Alternate textual formulations are used.	Not all data is trustworthy and heterogeneous. Data and large data spaces challenge resolution across many records	This is a big challenge. Statistical and machine learning techniques are used, but do we understand the range of possible errors that could occur in extracted facts?	(Zhu et al, 2020)

Final Graph Construction This step may include new links and confidences about facts and relations.	Working solution may not scale up to more data.	As more data is added, different vocabularies are introduced and different patterns may encode the same attribute.	Graph construction may be viewed as an incremental process with a final assembly that may include a check of semantic relations from a guiding ontology.	(Madison, Barnhill, Napier, and Godin, 2015; Deprizio, 2020)
User Interfaces	How flexible are the interfaces for the users? Can the KG be easily visualized when showing relations and entity linking?	Is explanation provided?	Allow customers to specify info and say which requirement is less important as part of query relaxation or refinement	(He et al, 2019)
Reasoning	How well does the KGA support temporal and spatial reasoning?		See the discussion in Section 4.3.	(Davis, 2020)

7. The Future of Knowledge Graphs

In this section we propose some possibilities for the future development and uses of KGs, primarily in industry but also for the KG research community.

1. There will be a general acceptance of an effective definition of "knowledge graph".

2. KG developers will understand the need for a well thought out schema and how ontologies, or at least ontological analysis, can aid in this.

3. KG developers will make use of linguistic analyses to help overcome the ambiguities of the use of natural language terms (and identifiers).

4. KG developers will incorporate formal distinctions for the intended interpretations of the natural language terms and phrases used for the labels in a KG, rather than the unfortunate practice of relying on assumed common interpretations for the semantics of such terms and phrases.

5. KGs will be used in the creation and operation of software intensive systems (e.g., representation of user interfaces).

6. Information systems architects will better exploit KGs and their infrastructure to support more dynamic information systems.

7. Architectures will be developed to aid enterprises and their extensive information systems in a transition to the use of KGs.

8. KGs will have a significant effect on data and knowledge management in general.

8. Conclusion

KGs are effective tools for information systems and are a very popular topic despite the lack of a common definition for what a KG is. This Communiqué has examined the notion of a KG and has made some progress toward specifying a succinct practical definition of what a KG is that not only is compatible with the main published definitions but also elucidates the sources of the confusion surrounding the notion of a KG. We have outlined the historical trends that converged on KGs and proposed some of the reasons why KGs have become so popular. Several examples of the techniques being used by KGAs and the KG systems that have been developed were described. Many standards now exist or are being developed that are relevant to KGs. While KGs have been successful, many issues and problems still remain.

9. Acknowledgments

Certain commercial software systems are identified in this paper. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology (NIST) or by the organizations of the authors or the endorsers of this Communiqué; nor does it imply that the products identified are necessarily the best available for the purpose. Further, any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NIST or any other supporting U.S. government or corporate organizations.

We wish to acknowledge the support of the ontology community, especially the invited speakers and participants who contributed to the Ontology Summit. There were 22 invited speakers: Jans Aasman, Andreas Blumauer, Barry Smith, Ernest Davis, Anirudh Prabhu, Sargur Srihari, Spencer Breiner, Paco Nathan, Vinay K. Chaudhri, Krzysztof Janowicz, Binil Starly, Sean Gordon, Michael Uschold, Yolanda Gil, Matthew West, Chaitanya Baru, Lisa Carnahan, Elisa Kendall, Michael Grüninger, John F. Sowa, Michael Bennett, and Elise Stickles. The complete list of sessions, speakers, and links to presentation slides and video recordings is available at http://bit.ly/33NydJ9.

A partial list of other contributors includes: Kingsley Idehen, Janet Singer, Doug Foxvog, Jack Hodges Jr, Alex Shkotin, Sjir Nijssen, Paul Tyson, Michael DeBellis, Edward Barkmeyer, Azamat Abdoullaev, Evan K. Wallace, Amit Sheth, Pascal Hitzler, Alessandro Oltramari, Jack Park, George Hurlburt, Russell Reinsch, Mariya Evtimova, and Bruce Bargmeyer. We especially thank Todd Schneider for suggesting the topic of knowledge graphs for the Ontology Summit 2020, and Kristina Rigopoulos for drawing Figure 4.

References

Aasman, J. (2019) Why Knowledge Graphs Hit the Hype Cycle and What they have in common. Retrieved on December 1, 2020 from http://bit.ly/34jSlmJ.

Aijal, J. (2019) What is a knowledge graph and how does one work? Retrieved on December 1, 2020 from http://bit.ly/2IwjVTu and https://thenextweb.com/podium/2019/06/11/what-is-a-knowledge-graph-and-how-does-one-work/.

Allen, R.H. & Sriram, D. (2000) The Role of Standards in Innovation, Special Issue on "Innovation: The Key to Progress in Technology and Society", Journal Technological Forecasting and Social Change.

Baclawski, K., Bennett, M., Berg-Cross, G., Casanave, C., Fritzsche, D., Ring, J., Schneider, T., Sharma, R., Singer, J., Sowa, J., Sriram, R.D., Westerinen, A. & Whitten, D. (2018) Ontology Summit 2018 Communiqué: Contexts in Context, J. Applied Ontology, IOS Press.

Baclawski, K., Futrelle, R., Fridman, N. & Pescitelli, M. (1993) Database techniques for biological materials & methods. In First Int. Conf. Intell. Sys. Molecular Biology 21-28.

Baclawski, K., Futrelle, R., Hafner, C., Pescitelli, M., Fridman, N., Li, B. & Zou, C. (1993) Data/knowledge bases for biological papers and techniques. In Proc. Sympos. Adv. Data Management for the Scientist and Engineer 23-28.

Baru, C. (2020) The Open Knowledge Network. Retrieved on December 1, 2020 from https://go.aws/31rSjbe.

Bennett, M. (2020) Standards for KGs in the Financial Sector. Retrieved on December 1, 2020 from https://go.aws/2YXCdXw.

Bergman, M. (2019) A common sense view of knowledge graphs. Retrieved on December 1, 2020 from http://bit.ly/307PEBs and http://bit.ly/2RAbE6X.

Blumauer, A. (2014) From Taxonomies over Ontologies to Knowledge Graphs. Retrieved on August 1, 2020 from https://blog.semantic-web.at/2014/07/15/from-taxonomies-over-ontologies-to-knowledge-graphs.

Breiner, S. (2020) Composing Knowledge Graphs, inside and out. Retrieved on December 1, 2020 from https://go.aws/2QfatbQ.

Carnahan, L. (2020) The IT Standard Process. Retrieved on December 1, 2020 from https://go.aws/3gPkyYK.

Chaudhri, V. (2020) Textbook Open Knowledge Network. Chaudhri Retrieved on December 1, 2020 from http://bit.ly/310xXpd.

Davis, E. (2020) Time and Space in Knowledge Graphs. Retrieved on December 1, 2020 from https://go.aws/2SxOVZ9.

Deprizio, J. (2020) Comparative Analysis of Database Spatial Technologies (CADST). Dissertation. George Mason University.

Dong, L. (2020) Knowledge Graph and Machine Learning: A Natural Synergy, Presentation at Stanford Seminar on KGs. Stanford University.

Dong, X., He, X., Kan, A., Li, X., Liang, Y., Ma, J., Xu, Y., Zhang, C., Zhao, T., Saldana, G., Deshpande, S., Manduca, A., Ren, J., Singh, S., Xiao, F., Chang, H.-S., Karamanolakis, G., Mao, Y., Wang, Y., Faloutsos, C., McCallum, A. & Han, J. (2020) AutoKnow: Self-driving knowledge collection for products of thousands of types, SigKDD 2020.

Färber, M., Ell, B., Menne, C., Rettinger, A., & Bartscherer, F. (2018) Linked Data Quality of DBPedia, Freebase, OpenCyc, Wikidata, and YAGO. Semantic Web Journal 9 (1), 77-129.

Gil, Y. (2020) Seven Ontologies for Publishing the Scientific Record on the Web. Retrieved on December 1, 2020 from https://go.aws/2yHUuO4.

Gordon, S. (2020) Prototyping an Open Knowledge Network for Spatial Decision Support. Retrieved on December 1, 2020 from http://bit.ly/2KquIjr.

Grüninger, M. (2020) Standards and Ontologies. Retrieved on December 1, 2020 from https://go.aws/2AykuMf.

He, X., Zhang, R., Rizvi, R., Vasilakes, J., Yang, X., Guo, Y., He, Z., Prosperi, M., Huo, J., Alpert, J. & Bian. J. (2019) ALOHA: developing an interactive graph-based visualization for dietary supplement knowledge graph through user-centered design. BMC medical informatics and decision making 19.4

"Industrial Ontology Foundry" (2020) Retrieved December 1, 2020 from https://www.industrialontologies.org/

Janowicz, K. (2020) KnowWhereGraph: Enriching and Linking Cross-Domain Knowledge Graphs using Spatially-Explicit AI Technologies to Address Pressing Challenges at the Human-Environment Nexus. Retrieved on December 1, 2020 from https://go.aws/2xmMSQd.

Joshi, P. (2019) Learn How to Perform Feature Extraction from Graphs using DeepWalk. Retrieved December 1, 2020 from https://www.analyticsvidhya.com/blog/2019/11/graph-feature-extraction-deepwalk/.

Kendall, E. (2020) The Object Management Group. Retrieved on December 1, 2020 from https://go.aws/3fISLc0.

Krötzsch, M. & Thost, V. (2016) Ontologies for knowledge graphs: Breaking the rules. In International Semantic Web Conference. Springer, Cham.

Liebmann, L. (2020) 3 reasons why DataOps is essential for big data success. In IBM Big Data & Analytics Hub. Retrieved October 28, 2020 from https://www.ibmbigdatahub.com/blog/3-reasons-why-dataops-essential-big-data-success.

Madison, M., Barnhill, M., Napier, C. & Godin, J. (2015) NoSQL database technologies. Journal of International Technology and Information Management 24.1.

Nathan, P. (2020) Rich Context Knowledge Graphs. Retrieved on December 1, 2020 from https://go.aws/2TwytYO.

Paulheim, H. (2017) Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. Semantic Web Journal 8(3):489-508.

Pham, M., Alse, S., Knoblock, C. & Szekely, P. (2016) Semantic labeling: a domain-independent approach. In International Semantic Web Conference. Springer, Cham.

Prabhu, A. (2020) Insights from Knowledge Graphs. Retrieved on December 1, 2020 from https://go.aws/3a9Niax.

Pujara, J., Miao, H., Getoor, L., and Cohen, W. (2013) Knowledge Graph Identification. In Proceedings of the 12th International Semantic Web Conference - Part I, ISWC'13, pages 542-557, New York, NY, USA.

Rohrseitz, N. (2019) Knowledge Graphs and Machine Learning: A powerful combination for the semi-automatic generation of insights. Retrieved on December 1, 2020 from https://towardsdatascience.com/knowledge-graphs-and-machine-learning-3939b504c7bc and http://bit.ly/2ZWVmqa.

"Semantic Network" (2020) Retrieved November 2, 2020 from https://bit.ly/36qXdct.

Smith, B. (2020) From BFO to IOF to ISO/IEC 21838. Retrieved on December 1, 2020 from https://go.aws/2zY2Otx.

Sowa, J. (2020) Knowledge Graphs and Logic. Retrieved on December 1, 2020 from https://go.aws/2LkvpeN.

Srihari, S. (2020) Probabilistic Knowledge Graphs. Retrieved on December 1, 2020 from http://bit.ly/36zrva9.

Starly, B. (2020) Building an Open Knowledge Network Graph in Product Design and Manufacturing. Retrieved on December 1, 2020 from https://go.aws/2Xna2Ay.

Taheriyan, M., Knoblock, C., Szekely, P. & Ambite, J. (2016) Leveraging Linked Data to Discover Semantic Relations Within Data Sources. In International Semantic Web Conference. Springer, Cham.

Wang, Y., Xu, Y., Li, X., Dong, X., Gao, J. (2020) Automatic validation of textual attribute values in eCommerce Catalog by learning with limited labeled data, In KDD’20.

West, M. (2020) The Digital Twin Project in the UK. Retrieved on December 1, 2020 from https://go.aws/2HdGBYr.

Yuan, W., Zhang, K., Dai, Q., Peng, C. & Zhao, K. (2018) Construction and Application of Knowledge Graph in Full-service Unified Data Center of Electric Power System. In IOP Conf. Ser.: Mater. Sci. Eng. 452 032065.

Zhu, Q., Wei, H., Sisman, B., Zheng, D., Faloutsos, C., Dong, X. & Han, J. (2020) Collective multi-type entity alignment between knowledge graphs. In WebConf 2020.