Published in final edited form as:

Annu Rev Biomed Data Sci. 2020 July ; 3: 23–41. doi:10.1146/annurev-biodatasci-010820-091627.

Knowledge-Based Biomedical Data Science

Tiffany J. Callahan[1], **Ignacio J. Tripodi**[2], **Harrison Pielke-Lombardo**[1], **Lawrence E. Hunter**[1]

[1]Computational Bioscience Program and Department of Pharmacology, University of Colorado Denver Anschutz Medical Campus, Aurora, Colorado 80045, USA

[2]Department of Computer Science, University of Colorado, Boulder, Colorado 80309, USA

Abstract

Knowledge-based biomedical data science involves the design and implementation of computer systems that act as if they knew about biomedicine. Such systems depend on formally represented knowledge in computer systems, often in the form of knowledge graphs. Here we survey recent progress in systems that use formally represented knowledge to address data science problems in both clinical and biological domains, as well as progress on approaches for creating knowledge graphs. Major themes include the relationships between knowledge graphs and machine learning, the use of natural language processing to construct knowledge graphs, and the expansion of novel knowledge-based approaches to clinical and biological domains.

Keywords

knowledge graph; ontology; natural language processing; knowledge discovery; Semantic Web; knowledge graph embeddings

INTRODUCTION

What Is Knowledge-Based Biomedical Data Science?

Knowledge-based biomedical data science (KBDS) involves the design and implementation of computer systems that act as if they knew about biomedicine.[1] There are many ways in which a system might act as if it knew something: For example, it might use existing knowledge to generate, rank, or evaluate hypotheses about a dataset, or it might answer a natural language question about a biomedical topic.

Knowledge-based systems have long been a theme in artificial intelligence research. Knowledge-based systems specify a knowledge representation—how a computer system represents knowledge internally—and one or more methods of inference or reasoning—how computations over those representations (perhaps combined with other inputs) are used to produce outputs. Classical descriptions of knowledge representation and reasoning systems

larry.hunter@cuanschutz.edu.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

[1]This is a revised and extended version of the introduction to Reference 1.

[e.g., see Davis et al. (2)] characterize them by the ontological commitments a knowledge representation makes (i.e., what it can or cannot describe), which inferences are possible within it, and, sometimes, which of those inferences can be made efficiently. These issues remain useful in thinking about how knowledge representation and reasoning play a role in today's data science environment.

This review provides some useful background knowledge on important KBDS concepts like ontologies, Semantic Web standards, and the distinction between knowledge bases and knowledge graphs (KGs). To provide context, we then describe some high-level applications of KBDS that were published prior to January 2020. Then, we describe our approach to reviewing the last year of KBDS research and present our findings. Finally, we conclude the review by discussing perceived barriers to and offering recommendations for conducting KBDS.

Ontologies

Ontologies are a vital component of knowledge representations. Knowledge representations are said to be grounded in a set of primitive terms, hereafter termed "primitives," that specify those ontological commitments: the entities and processes that can be referred to by that knowledge representation. Computational ontologies are, then, collections of primitives relevant to a domain, often related to each other by explicit subsumption (subclass of) and meronomy (part of) statements.

Within the biomedical domain, ontologies [e.g., the Gene Ontology (GO) (3)] are community consensus views of the entities involved in biology, medicine, and biomedical research, analogous to how nomenclature committees systematize naming conventions. Knowledge bases created using primitives from community-curated ontologies, rather than idiosyncratic or single-use sets of primitives, provide significant advantages for reproducibility in scientific research, for interoperability, and for avoiding pitfalls in the modeling of knowledge. Primitives can be combined into assertions that express facts about the world. In the simplest case, an assertion links two primitives with a specific relationship. Consider, for example, that the Protein Ontology contains a primitive human p53 protein, the GO contains a primitive DNA strand renaturation, and the Relation Ontology contains a primitive participation that can link a physical entity to a process in which it participates. Those three primitives can be composed into an assertion that could be part of a knowledge base: Human p53 protein participates in DNA strand renaturation. Some KGs are grounded in terminological resources, such as UMLS (Unified Medical Language System), SNOMED CT (Systematized Nomenclature of Medicine, Clinical Terms), and the National Cancer Institute Thesaurus, that lack some aspects of a computational ontology.

Semantic Web Standards

While ontologies provide the primitive elements from which a knowledge representation is constructed, they are agnostic about the mechanisms by which entities are assembled into assertions. In 2011, the World Wide Web Consortium promulgated a collection of international standards for linking entities with shared meaning into assertions and managing collections of assertions, together termed the Semantic Web (https://www.w3.org/standards/

semanticweb/). The Semantic Web builds on the standard Resource Description Framework (RDF) (https://www.w3.org/standards/techs/rdf), which provides a way to link three uniform resource identifiers (4) to specify a relationship between a pair of entities (forming an RDF triple). Collections of triples form a graph, where the entities are nodes and the relationships are edges connecting them. A computational mechanism for managing such collections is called a triplestore.

The Semantic Web standards also define RDF Schemas and a Web Ontology Language (OWL), which provide additional expressivity; SPARQL [SPARQL Protocol and RDF Query Language (5)], which provides a query language for interrogating RDF graphs or triplestores; and the Simple Knowledge Organization System, which provides a basic ontology. The OWL (6) is used to specify two important types of entities: instances and classes. Instances are particular entities or processes in the world (e.g., a particular molecule of p53) and classes are groupings of instances that meet a defined set of individually necessary and collectively sufficient criteria (e.g., human p53 proteins). As it lacks variables and quantification, OWL cannot express all logical statements about primitives; the subset of first-order logic that OWL can express is inspired by description logics (7).

Knowledge Base Versus Knowledge Graph

Collections of assertions, generally called knowledge bases, can be created, queried, and shared, and then in turn used by other systems that apply various inference methods to fulfill particular application needs. Knowledge bases that can be represented as graphs are often called “knowledge graphs.” While not all knowledge bases are implemented as graphs (e.g., some are databases where a table structure is used to make implicit assertions), in recent years, it has become very common to represent knowledge bases using Semantic Web standards, or at least to use knowledge bases that can produce and consume Semantic Web–compatible versions (8). For that reason, the terms “knowledge base” and “knowledge graph” are often used interchangeably. In 2012, Google announced its proprietary Knowledge Graph, which also popularized the use of the term (9). The literature sometimes contains terminological imprecision about the differences between knowledge bases, KGs, and ontologies; readers are referred to Reference 10 for a recent review and analysis of various published definitions.

In this review, we use the term “knowledge graph” (“KG”) and say that a KG is grounded in the set of primitives from which it is constructed. Some KGs also include a set of logical rules called axioms that relate assertions to each other (e.g., the human p53 protein is a subclass of the p53 protein class that is found in the organism human). Figure 1 shows a simple example of a biomedical KG.

Biomedical Applications of Knowledge Graphs

KBDS does computation over KGs (and perhaps other inputs) to make inferences about biomedicine. While each of the works surveyed below addresses a different problem using a different technique, there are some common themes in the computational approaches to using KGs, including improving information retrieval, inferring new knowledge, and creating alternative representations of KGs. Each of these themes is discussed further below.

Information retrieval.—A major use of KGs is simply to organize knowledge for information retrieval. Such systems are designed to make it possible to find facts or evidence regarding a wide variety of topics, ranging in this review from cataloging traditional Chinese medical practices to decision support for pharmacovigilance. KGs have also been used to improve other forms of information retrieval, such as finding relevant publications in the literature.

Inferring new knowledge.—There are two primary ways that new information in KGs is automatically inferred: graph algorithms and logical reasoning.

Edge (or link) prediction is one class of graph algorithm that is widely used in KBDS. Edge prediction methods generally use the structure of a graph to identify edges that are likely but missing in the graph (11). In KGs, these are predictions of assertions about the world. This is a form of hypothesis generation and often includes an estimate of the confidence in the prediction. Many approaches to drug repurposing use edge prediction algorithms over KGs of drugs and diseases to identify new indications. Another broad class of graph algorithms does community finding, or identification of groups of entities in a KG that are similar or highly related to each other. For example, some approaches to disease subphenotyping apply community-finding approaches to KGs encoding information about patients.

The second way new information is inferred from KGs is through the use of logical reasoners. The Semantic Web OWL standard was designed to facilitate two important classes of reasoning over KGs: satisfiability and subsumption inference. Satisfiability inference checks to see if a class definition is logically satisfiable; it is possible for a KG to define a class that has no members (e.g., human p53 protein homologs in bacteria). Subsumption inference uses class definitions to identify all classes that are fully contained within some other class (e.g., all proteins are nitrogen-containing compounds). Specific reasoners, such as ELK (12) or HermiT (13), can be used to make these inferences with particular computational performance guarantees, which can be important in large KGs. Subsumption inference in particular is useful in KBDS because it makes explicit many edges that are otherwise implicit in KGs, and therefore it can improve the results of other algorithms that depend on the structure of the graph, such as link prediction or embeddings.

Alternative representations.—Machine learning, particularly in the form of artificial neural networks, is widely used in the KG context (14). One frequent application of neural networks to KGs is to create embeddings of entities or assertions by training autoencoder networks with inputs constructed from the KG. These embeddings can then be used to compute knowledge-based similarities between, e.g., drugs, proteins, and diseases. Neural network methods have also been used to identify parts of a KG relevant to question and answering (Q&A) for a given input question.

Known Challenges

Many challenges to designing, constructing, and utilizing KGs within the biomedical domain have been identified. Some of the more difficult challenges include (*a*) computational performance, (*b*) constructing KGs using expert curation or information

extraction methods, and (*c*) meaningfully integrating disparate data sources. Each of these challenges is discussed further below.

Computational performance.—Computational performance is always a challenge when applying algorithmically complex methods to large volumes of data. Biomedical knowledge is very extensive, and broad biomedical KGs can contain billions of assertions. A wide variety of schemes have been proposed to address the computational complexity of both querying and inference over KGs (15).

Knowledge graph construction.—A few KGs, such as GO annotations (16), GO Causal Activity Models (GO-CAMs) (17), or Reactome (18), are constructed through intense expert curation efforts. However, several algorithmic approaches have been proposed to either augment these efforts or fully automate them. Automated approaches to KG construction fall into two broad classes: natural language processing (NLP) methods and data-driven construction. Data-driven KG construction can involve the integration of previously disparate resources or the direct analysis of large-scale datasets.

NLP methods propose to extract information from a set of documents to create KGs, for example, the Semantic MEDLINE Database (SemMedDB) (19). As NLP methods are all imperfect, these approaches are often focused on assessing the reliability of the information extracted, or on techniques to manage missing or erroneous assertions and other sources of noise.

Data integration.—Some data-driven approaches simply transform existing databases [e.g., DrugBank (20)] into KGs, which can be useful for tasks like facilitating adherence to FAIR (findable, accessible, interoperable, and reusable) research principles (21). More frequently, data-driven KG construction aims to integrate multiple sources of data into a single KG. If an integrated KG can ground the different sources in one set of primitives (ideally, from a community-curated ontology), then inference over the combined information can be facilitated. There are thousands of public, biomedically important databases (22); hence, integration approaches that support semantic compatibility are important and can lead to improved data quality, as incompatibilities sometimes signal errors (23).

METHODS OF REVIEW PROCESS

For this review, relevant literature was surveyed by searching PubMed and Google Scholar using the following phrases: “knowledge graph,” “biomedicine”; “knowledge graph,” “medicine”; “knowledge graph,” “medical”; and “knowledge graph,” “biology.” These terms were also searched by replacing “graph” with “base.” All paper types were eligible for inclusion (i.e., conference proceedings, dissertations, book chapters, peer-reviewed archived manuscripts, and published peer-reviewed manuscripts).

Results

The search phrases above returned 52 papers from PubMed and 7,752 papers from Google Scholar. Manual review of these papers was performed to identify those that were focused on the use or construction of KGs within the biomedical domain, which resulted in a

reduced set of 174 papers. This set of papers was then further reduced to only include papers published or posted to public manuscript archives between January 2018 and December 2019 whose full-text version was publicly available at the time of writing, resulting in a final set of 83 papers. For additional details, please see Figure 2.

The final set of papers was further broken down by year of publication or presentation/submission, the manuscript type, and the journal or archive name. Among the 83 papers, 41 were published or posted online in 2018. The majority of 2018/2019 papers were published in conference proceedings ($n = 38$) or in peer-reviewed journals/books ($n = 26$), with the remaining papers posted as online preprints ($n = 19$). Among these, the majority of the 2018 and 2019 papers were published as conference proceedings (48.8% and 42.9%, respectively). The number of conference submissions decreased by 9.1% between 2018 and 2019; 2018 papers were primarily submitted to IEEE (Institute of Electrical and Electronics Engineers) ($n = 10$), whereas 2019 papers were submitted to ACM (Association for Computing Machinery) ($n = 5$), AAAI (Association for the Advancement of Artificial Intelligence) ($n = 3$), and IEEE ($n = 3$).

Organization and Presentation of Findings

The final set of papers fall into two broad categories, which are used to organize the remainder of this review: application of KGs ($n = 53$) versus construction of KGs ($n = 30$). This review then concludes by considering some nascent projects likely to be important in the near future, characterizing current barriers to building and using biomedical KGs, and making some recommendations. Information about each paper included in the final set is presented in Supplemental Tables 1 and 2, and broad themes spanning multiple papers are described below.

APPLICATIONS OF KNOWLEDGE GRAPHS IN BIOMEDICAL DATA SCIENCE

Applications of KGs are noted in a wide variety of biomedical domains, ranging from analysis of genomic data to clinical decision support. There is also a close relationship between KGs and biomedical NLP: KGs can be used to improve the quality of NLP, and NLP can be used to generate KGs from the literature. Thus, the application of KGs is further divided into three primary themes: (*a*) clinical, (*b*) biological, and (*c*) NLP. Supplemental Table 1 provides a high-level summary of the reviewed papers that used KGs to help solve a biomedical data science problem (References 24-79).

Clinical Applications

There were three primary themes identified within this domain in the papers surveyed, including the use of KGs to improve the retrieval of information from the literature or from large sources of clinical data, the use of KGs to provide confidence either by adding evidence to support phenomena observed in data or by completing missing information and deriving new hypotheses, and the use of KGs to improve the representation and presentation of complicated patient data or personal health information.

The first observed theme was the use of KGs to refine user queries and otherwise improve information retrieval from the literature or from an electronic health record (EHR) system.

One study demonstrated that using KGs with traditional rule-based approaches for information retrieval performed better than using either KGs or rule-based approaches alone (24). Liu et al. (25) proposed a novel graph-based representation of patient data where entities were linked to concepts in a biomedical KG in order to enable querying based on domain knowledge. Other clinical applications of KGs in information retrieval included the finding that a KG-based component added to a larger system improved the ability of doctors to identify meaningful information from an EHR (26), a KG-based method for users to formulate queries within the context of relevant domain knowledge (27), and a system to rewrite user queries using domain knowledge (28).

The second observed theme was the use of KGs to address uncertainty by identifying relevant evidence. KGs have been leveraged to provide evidence for diagnostic assistance, clinical decision support machinery, or surveillance (29). For example, Bakal et al. (30) used SemMedDB and a subset of the UMLS to better predict treatments for and causes of different diseases, and Reumann et al. (31) found that using a KG was helpful for correctly identifying rare disease patients when examined using over 100 different queries. There were two papers that focused on surveillance. In the first paper, Bobed et al. (32) built a KG from an adverse drug reaction ontology and SNOMED CT as a means to improve pharmacovigilance. In the second paper, Kamdar et al. (33) built a KG using drug classes from the Anatomical Therapeutic Chemical Classification System and active ingredients in RxNorm to better understand opioid use patterns across the United States.

Also part of the second observed theme was the use of link prediction algorithms to discover missing knowledge or generate hypotheses (e.g., 25). To improve the identification of comorbid diseases, Biswas et al. (34) built a KG using the approach outlined by Alshahrani et al. (35) and then performed link prediction using an inductive inference method. Also using inductive inference methods, Callahan et al. (36) described a method for transforming OWL-encoded knowledge to create representations that were better suited to inductive inference tasks; the results were evaluated using queries against the Knowledge Base of Biomedicine (KaBOB) (37). Neil et al. (38) described a method for transforming KGs into graph convolutional neural networks and an attention model using independent learnable weights to measure each edge’s usefulness.

The final observed theme was the use of KGs to capture complex patient information for further processing. Xie and colleagues (39) created patient-specific traditional Chinese medicine KGs by mapping patient data to a general traditional Chinese medicine–specific KG. Shang et al. (40) described a method for creating visit-level representations of patients from EHR data and mapping to a drug–drug interaction KG to provide personalized medication combination recommendations. Three papers focused on how to improve the presentations of complex information or results. Huang et al. (41) developed a novel tool to enrich and visualize patient data by incorporating KG embeddings. Singh et al. (42) developed an interactive tool built on Cytoscape (43) to help users interact with their network data. And Queralt-Rosinach et al. (44) introduced a novel approach to create custom systematic literature reviews by formulating the review as a biomedical KG that contains information relevant to specific hypotheses provided by a user.

Biological Applications

In more basic research applications, broad themes included the use of KGs to produce embeddings for prediction or visualization in low-dimensional spaces (15, 45, 46), the use of link prediction methods over KGs to hypothesize previously unobserved relationships (36, 40, 45, 47-56), and the use of KGs to generate complex mechanistic accounts of experimental data. Several efforts combined these themes, particularly the use of edge embeddings to improve link prediction (34, 48, 53, 57, 58).

Node and edge embeddings provide a powerful method to suggest relationships among entities via similarity functions, in ways that complement path traversal through the graph. Many of the reviewed papers leveraged semantic similarity–inspired hypotheses to identify valuable drug–drug (40,47, 51), drug–target (51), or protein–protein interactions (46, 48), many of which were in turn applied to drug repurposing. Additionally, some of the papers converted KG-based embeddings into low-dimensional spaces in order to visualize clusters in two- or three-dimensional projections (41) to better display entities of interest.

In a particularly innovative approach, Tripodi et al. (45) combined gene expression time series and KG-based embeddings from a human-centric KG (59) to create specific and detailed hypotheses regarding mechanisms of toxicity. The resulting KG subgraphs that made up the hypothesized mechanisms were far richer than the black box toxicity predictions that would have been otherwise presented. Further, these subgraphs were also used to generate natural language narratives describing the mechanisms and their sources of evidence.

Natural Language Processing Applications

KGs have been used to improve NLP performance in a wide variety of genres, including summarization or information extraction from EHRs and Q&A systems (15, 26, 27, 29, 40, 60, 61). We observed that KG-derived embeddings used alone or in combination with other text-derived features (46) improved the performance of a variety of NLP tasks, including named entity recognition (62), coreference resolution (63), and relation extraction (64).

Additionally, several papers demonstrated the utility of KGs in information extraction methods. Ontologies can serve as formal dictionaries allowing for rapid indexing in named entity recognition and word sense disambiguation tasks (65, 66). Compared to lexicons, KGs offer far richer semantic context, identifying not only similar concepts but also rich collections of relationships that can be used to disambiguate or otherwise improve concept recognition in texts (65-68).

CONSTRUCTING KNOWLEDGE GRAPHS

Researchers have made substantial efforts in methods development and generated new results in the construction of KGs, as well as in extending, integrating, and evaluating them. Supplemental Table 2 provides a high-level summary of the papers surveyed on the construction of KGs (References 80-124).

Efforts to produce domain-specific KGs have been made in a variety of areas, including biodiversity (80-82) and the microbiome (https://ncats.nih.gov/translator), as well as for the purpose of enriching clinical data (83, 84). The papers surveyed on biodiversity focused specifically on how a KG could be created and linked to identifiers in the literature (80, 81) or other important biodiversity resources (82). In contrast, papers using KGs for clinical enrichment aimed to use them as a way to link clinical data to sources of evidence to provide support for clinical observations (84-86) or to help make the data more interpretable with respect to underlying biological mechanism(s) (83) for improved diagnosis (87).

Historically, NLP information extraction efforts have often been used to construct KGs; two novel methods to do so were published last year. One proposed a minimum supervision-based approach that combined traditional NLP pipelines for information extraction and biomedical context embeddings (88). The other focused on improving the extraction of biomedical facts from the literature by leveraging and refining specific seed patterns (80).

Although not a construction method per se, SemMedDB, one of the most widely used NLP-constructed KGs, was recently evaluated by Cong et al. (81). They found many contradictory assertions in a variety of fundamental relationship categories, underscoring the need to be cautious regarding noise in NLP-derived KGs. Finally, an ontology called BioKNO and a set of associated tools leveraging OWL were presented to assist scientists attempting to share data according to FAIR principles (82).

ORGANIZATIONAL EFFORTS IN KNOWLEDGE-BASED BIOMEDICAL DATA SCIENCE

Both US and European scientific institutions support KG efforts. Perhaps the most ambitious of these is the National Institutes of Health's National Center for Advancing Translational Sciences' Biomedical Data Translator project (https://ncats.nih.gov/translator). The goal of the Translator is a computational system that integrates sources of existing biomedical knowledge in order to translate clinical inquiries into relevant biomedical research results that synthesize elements of the integrated knowledge to directly answer the inquiry or generate testable hypotheses (83). A recent funding call targets $13.5 million per year for up to five years toward the construction of what they call "knowledge providers and autonomous relay agents" (89). Knowledge providers are systems that seek out, integrate, and provide high-value data sources within a specific scope of Translator-relevant knowledge, and presumably they would primarily use KGs to do so. Relay agents are to take clinical queries in a standardized format, dispatch subtasks to appropriate knowledge providers, receive responses back from knowledge sources (presumably also as subgraphs of a KG), and process responses using scoring metrics in order to return the most relevant and highest quality potential responses.

Investigators at the University of California San Francisco (UCSF), Google, the Lawrence Livermore National Laboratory, and the Institute for Systems Biology were recently awarded the National Science Foundation's Convergence Accelerator Award (84). A total of 21 awards were given out, but the UCSF project was the only one that focused on solving biomedical and health-related problems. The awarded project, titled "A Multi-Scale Open

Knowledge Network for Precision Medicine," aims to integrate several sources of publicly available data in order to build what they term a "biomedical knowledge engine." The long-term goal of this project is to create a resource that will help clinicians gain better insight into patient care, as well as provide a tool to aid biomedical research. The project will be developed using UCSF's Scalable Precision Medicine Knowledge Engine (SPOKE) (85, 86). By incorporating additional data from the UCSF Information Commons (https://informationcommons.ucsf.edu/), SPOKE will extend Hetionet (87), which currently contains information from over 25 databases and links millions of drugs, diseases, and biological molecules. In collaboration with Google, SPOKE will eventually be made available to the public through Data Commons (90).

Elixir Europe (https://elixir-europe.org/) is a large multinational (and European Commission) project with the goal of managing and safeguarding the data generated by publicly funded life science research and integrating bioinformatics resources. In pursuit of those goals, Elixir's interoperability platform promotes efforts in the European life science community to adopt standardized file formats, metadata, vocabulary, and identifiers, including work on the Semantic Web and the adoption of community-curated ontologies. The Elixir Core Data Resources (https://elixir-europe.org/services/tag/core-data-resources) are leaders in the production of interoperable knowledge resources and are widely used components of biomedical KGs.

Last year saw the announcement of GO-CAMs, a new approach to GO annotation (17). Although GO annotations are perhaps the most widely used knowledge representations in biomedical research, until GO-CAMs were introduced, the annotations could not be assembled into a coherent KG. While individual annotations implicitly linked GO classes to gene products, contextual information was lost: For example, the annotation process could not capture that cytochrome C participated in apoptosis only when it was in the cytoplasm. GO-CAM models and associated tooling are gradually replacing the traditional GO annotation process within the Alliance for Genomic Resources (81), meaning that future GO annotation will produce an increasingly rich, manually curated KG.

CONCLUSIONS AND RECOMMENDATIONS

As demonstrated by this review, the last year has seen tremendous amounts of new developments in both the construction and application of biomedical KGs. A significant number of the reviewed papers focused on the construction or application of KGs to solve problems within the clinical domain (e.g., providing evidence for traditional Chinese medicine, improving Q&A systems, and developing patient-level KGs). Another popular area observed in both biological and NLP-based applications of KGs was the development of novel methods to better utilize KGs (e.g., embedding algorithms to create alternative representations of data extracted from a KG and graph-based algorithms to improve information and relation extraction). Finally, we observed across all applications of KGs that while KGs provide an efficient way to present complex information (e.g., scientific and medical knowledge, biological interactions, and experimental results), user-friendly tools are still needed to help visualize and present this information. We also identified several challenges and barriers to construct and use KGs that emerged from these papers (and also

from our shared research experiences) that have helped us to brainstorm solutions and recommendations. Each of these areas is further detailed below.

Barriers

Current barriers to constructing and using KGs include KG and data availability, data licensing issues (sometimes there are different licenses for each data source), a lack of agreed upon standards for constructing KGs, and dependency upon resources (i.e., software languages or applications) that may be obsolete, deprecated, or outside of users' skill sets or areas of expertise.

The first barrier to using KGs is that building a KG is challenging, so the reuse of existing resources is highly desirable. Table 1 provides a list of currently available KGs. While each of these KGs is a valuable resource, each was developed to solve a specific problem and thus there may be challenges in applying it to new tasks. GO-CAMs have great promise, but currently only a relatively small number have been curated (17). Reactome (18) provides a very high quality and extensive KG grounded in community-curated ontologies, but it is limited in scope to biochemical reactions and pathways. KGs derived from manual or automatic data integration, such as KaBOB (37), PheKnowLator (59), Hetionet (87), SPOKE (85), Bio2RDF (91, 92), DisGeNET (93), BioGrakn (94), and the Data Commons Graph (90), all require different amounts of domain knowledge or technical expertise to utilize. KaBOB is grounded in community-curated ontologies, but licensing restrictions mean that users must download software and build the KG themselves, which requires expertise and computational resources. PheKnowLator is also grounded in community-curated ontologies, is deductively closed using ELK, and is publicly available, but it does not yet have a user interface. Bio2RDF, BioGrakn, the Data Commons Graph, DisGeNET, Hetionet, and SPOKE are constructed by combining several different types of data without a consistent set of primitives and are not fully grounded in ontologies. Finally, NLP-derived systems such as SemMedDB are noisy, making trustworthiness an issue. This review does not discuss all of the currently available biomedical KGs identified during our literature survey. Readers are referred to Table 1 for additional information.

KGs constructed from automatic methods also present significant barriers. KG construction from literature sources is usually framed as a relation extraction problem, where semantic triples are inferred from the text and then assembled into a KG. The correctness of this approach to KG construction can be determined either before the KG is constructed by evaluating the relation extraction process itself (a more traditional approach) or by evaluating the quality of the resulting KG itself. Evaluating the quality of the constructed KG allows for the use of the reasoning methods described above.

The final barrier is the lack of agreed upon standards for evaluating KGs. In fact, the lack of standards for constructing KGs within the biomedical domain may be one of the reasons why they are challenging to evaluate. Of the 25 reviewed papers on constructing KGs that evaluated their KGs, 4 provided qualitative evaluation (e.g., case studies, domain expert or focus group review of results, conceptual models, or prototypes), 5 provided quantitative evaluation (e.g., by applying machine learning to a specific holdout dataset or to a new dataset, by performing a KG completion task like edge prediction), and 16 provided both

types of evaluation. One of the reviewed papers that provided both types of evaluation utilized crowdsourcing as a means to validate triples from their KG (95).

Recommendations

Based on the articles surveyed, we provide a brief list of recommendations below. Two of the most significant areas that deserve attention are the use of dense vector representations (embeddings) of concepts and the integration of KGs in NLP applications.

Knowledge graph embeddings.—Given the high volume of papers covering KG embeddings and NLP-based KGs, we felt it necessary to share recommendations gleaned from these works. A very active area of research is the use of KGs to create knowledge-based embeddings. Good embeddings are important to the performance of machine learning systems and therefore have wide applicability. KGs have been used to create embeddings for entities of many kinds (from genes to patients), as well as for relations, assertions, and more complex representations. Applications of these embeddings include prediction of drug–drug interactions, drug–target interactions, target discovery, and finding clinically relevant evidence. In addition to reusing embeddings from the surveyed papers, we recommend considering the tools described in the BioKEEN paper (96), which describes a Python-based library for training and tuning models to produce new knowledge-based embeddings.

Natural language processing-based knowledge graphs.—As previously described, a major theme of the reviewed literature is the use of text mining and NLP techniques to generate KGs. While this approach offers the potential for breadth missing from most manually curated KGs, it comes at the cost of a relatively large number of errors. Cong et al.'s (81) evaluation of SemMedDB, a widely used KG produced by the US National Library of Medicine, found nearly 500,000 inconsistent assertions, as well as a wide variety of apparently missing relationships. While they suggested methods that could be used to improve the quality of SemMedDB, our recommendation is to recognize that NLP-generated KGs are likely to be very noisy and need to be used with caution.

Future Work

The trends we observed in last year's work are likely to continue. Applications of KGs will likely continue to involve generations of embeddings and other uses of KGs in machine learning. The close relationships between the development of NLP methods and KGs are likely to persist. The expansion of KGs to areas beyond molecular biology (e.g., biodiversity and traditional Chinese medicine this year) is also likely to continue. Some previous areas of research (e.g., KG-based enrichment analysis for gene sets) that did not see new results this year may also continue to be fruitful.

New methods applying KGs to analyze different sorts of experimental data such as images seem ripe for development. Robust and biologically meaningful ways to incorporate or add experimental data to biomedical KGs would help to improve the precision of predictions when used to generate novel hypotheses or as a means to interpret experimental results. Similarly, for clinical KGs, it will be important to find clinically meaningful ways to

incorporate quantitative measures (e.g., laboratory test results and biomarker measurements) and outcomes from EHR data.

Other than Table 1 of this review, there is no central reference site or repository where one can access or identify a current list of available biomedical KGs. Researchers need a more systematic approach to the development, maintenance, and interoperability of biomedical KGs that facilitates sharing and the use of clear documentation. Existing efforts on general frameworks and tools like BioKEEN (96), Protégé (97), and BlazeGraph (98), in addition to open biomedical ontologies (99), are important and could be further extended toward standardized tool development. As KG evaluation remains challenging, new methods or benchmarks will be valuable.

A final area for future work is the development of tools for interacting with KGs. Protégé (97) and SPARQL are two of the most commonly utilized tools within the field, but both have limitations that make them unable to serve as comprehensive tool suites. While Protégé is useful for constructing and editing ontologies and for performing logical reasoning, it was not designed to efficiently handle very large KGs and has limited support for visualizing ontologies and KGs. SPARQL's limitations in pathway search (100) and ease of use make its broader adoption challenging. Recently developed applications like the Data Commons Graph web browser (90), ROBOKOP (61), and Grakn Enterprise's Knowledge Graph Management System and Workbase (https://grakn.ai/grakn-kgms) provide promising examples of well-crafted, sustainable user interfaces, but all of these applications are written in different programming languages and their proper use requires differing levels of programming expertise.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

LITERATURE CITED

1. Hunter LE. 2017. Knowledge-based biomedical data science. Data Sci. 1(1-2):19–25
2. Davis R, Shrobe H, Szolovits P. 1993. What is a knowledge representation? AIMag. 14(1):17–33
3. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. 2000. Gene Ontology: tool for the unifi-cation of biology. Nat. Genet 25(1):25–29 [PubMed: 10802651]
4. Berners-Lee T, Fielding RT, Masinter L. 2005. Uniform resource identifier (URI): generic syntax. Unpub-lished Memo., Internet Eng. Task Force, Fremont, CA. https://tools.ietf.org/html/rfc3986
5. SPARQL (SPARQL Protoc. RDF Query Lang.) Work. Group. 2013. SPARQL 1.1 protocol. Web Resour., World Wide Web Consort. https://www.w3.org/TR/sparql11-protocol/
6. W3C (World Wide Web Consort.). 2004. OWL Web Ontology Language overview. Web Resour., World Wide Web Consort. https://www.w3.org/TR/owl-features/
7. Krötzsch M, Simancik F, Horrocks I. 2012. A description logic primer. arXiv:1201.4089 [cs.AI]
8. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, et al. 2007. Advancing translational research with the Semantic Web. BMC Bioinform. 8(Suppl. 3):S2
9. Singhal A. 2012. Introducing the Knowledge Graph: things, not strings. Google Blog, 5 16. https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html
10. Ehrlinger L, Wöß W. 2016. Towards a definition of knowledge graphs. In Proceedings of the Posters and Demos Track of 12th International Conference on Semantic Systems (2016 SEMANTiCS). CEUR Workshop Proc.

11. Lü L, Zhou T. 2011. Link prediction in complex networks: a survey. Physica A 390(6):1150–70
12. Kazakov Y, Krötzsch M, Simancik F. 2012. ELK reasoner: architecture and evaluation. In Proceedings of the OWL Reasoner Evaluation Workshop (ORE 2012), ed. Horrocks I, Yatskevich M, Jiménez-Ruiz E. CEUR Workshop Proc.
13. Glimm B, Horrocks I, Motik B, Stoilos G, Wang Z. 2014. HermiT: anOWL 2 reasoner. J.Autom. Reason 53(3):245–69
14. Nickel M, Murphy K, Tresp V, Gabrilovich E. 2016. A review of relational machine learning for knowledge graphs. Proc. IEEE 104(1):11–33
15. Wang X, Wang Y, Gao C, Lin K, Li Y. 2018. Automatic diagnosis with efficient medical case searching based on evolving graphs. IEEE Access 6:53307–18
16. Gene Ontol. Consort. 2009. Introduction to GO annotations. Web Resource, Gene Ontol. Consort. http://geneontology.org/docs/go-annotations/
17. Thomas PD, Hill DP, Mi H, Osumi-Sutherland D, Van Auken K, et al. 2019. Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. Nat. Genet 51:1429–33 [PubMed: 31548717]
18. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, et al. 2018. The Reactome pathway Knowledgebase. Nucleic Acids Res. 46(D1):D649–55 [PubMed: 29145629]
19. Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. 2012. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics 28(23):3158–60 [PubMed: 23044550]
20. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, et al. 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 34(Suppl. 1):D668–72 [PubMed: 16381955]
21. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 [PubMed: 26978244]
22. Rigden DJ, Fernandez XM. 2019. The 26th annual Nucleic Acids Research database issue and Molecular Biology Database Collection. Nucleic Acids Res. 47(D1):D1–7 [PubMed: 30626175]
23. Cormont S, Vandenbussche P-Y, Buemi A, Delahousse J, Lepage E, Charlet J. 2011. Implementation of a platform dedicated to the biomedical analysis terminologies management. AMIA Annu. Symp. Proc 2011:1418–27 [PubMed: 22195205]
24. Chen Q, Li B. 2018. Retrieval method of electronic medical records based on rules and knowledge graph. In Proceedings of the 2018 International Conference on Electronic Business, Art. 42. Atlanta, GA: Assoc. Inform. Syst.
25. Liu X, Jin J, Wang Q, Ruan T, Zhou Y, et al. 2018. PatientEG dataset: bringing event graph model with temporal relations to electronic medical records. arXiv:1812.09905 [cs.CY]
26. Liu Z, Peng E, Yan S, Li G, Hao T. 2018. T-Know: a knowledge graph-based question answering and information retrieval system for traditional Chinese medicine. In Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pp. 15–19. New York: Assoc. Comput. Linguist.
27. Ruan T, Huang Y, Liu X, Xia Y, Gao J. 2019. QAnalysis: a question-answer driven analytic tool on knowledge graphs for leveraging electronic medical records for clinical research. BMC Med. Inform. Decis. Making 19(1):82
28. Mohammadhassanzadeh H, Abidi SR, Van Woensel W, Abidi SSR. 2018. Investigating plausible reasoning over knowledge graphs for semantics-based health data analytics. In Proceedings of the IEEE 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises, pp. 148–53. Los Alamitos, CA: IEEE Comput. Soc.
29. Schwertner MA, Rigo SJ, Araújo DA, Silva AB, Eskofier B. 2019. Fostering natural language question answering over knowledge bases in oncology EHR. In Proceedings of the IEEE 32nd International Symposium on Computer-Based Medical Systems, pp. 501–6. New York: IEEE
30. Bakal G, Talari P, Kakani EV, Kavuluru R. 2018. Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations. J. Biomed. Inform 82:189–99 [PubMed: 29763706]

31. Reumann M, Giovannini A, Nadworny B, Auer C, Girardi I, Marchiori C. 2018. Cognitive DDx assistant in rare diseases. In Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 3244–47. New York: IEEE
32. Bobed C, Douze L, Ferré S, Marcilly R. 2018. Sparklis over PEGASE knowledge graph: a new tool for pharmacovigilance. In Proceedings of the 2018 International Conference on Semantic Web Applications and Tools for Life Sciences (SWAT4LS), ed. Baker CJO, Waagmeester A, Splendiani A, Beyan AD, Marshall MS. CEUR Workshop Proc.
33. Kamdar MR, Hamamsy T, Shelton S, Vala A, Eftimov T, et al. 2019. A knowledge graph-based approach for exploring the U.S. opioid epidemic. arXiv:1905.11513 [cs.CY]
34. Biswas S, Mitra P, Rao KS. 2019. Relation prediction of co-morbid diseases using knowledge graph completion. IEEE/ACM Trans. Comput. Biol. Bioinform In press
35. Alshahrani M, Khan MA, Maddouri O, Kinjo AR, Queralt-Rosinach N, Hoehndorf R. 2017. Neuro-symbolic representation learning on biological knowledge graphs. Bioinformatics 33(17):2723–30 [PubMed: 28449114]
36. Callahan TJ, Baumgartner WA, Bada M, Stefanski AL, Tripodi I, et al. 2017. OWL-NETS: transforming OWL representations for improved network inference. Proc. Pac. Symp. Biocomput 23:133–44
37. Livingston KM, Bada M, Baumgartner WA Jr., Hunter LE. 2015. KaBOB: ontology-based semantic integration of biomedical databases. BMC Bioinform. 16:126
38. Neil D, Briody J, Lacoste A, Sim A, Creed P, Saffari A. 2018. Interpretable graph convolutional neural networks for inference on noisy knowledge graphs. arXiv:1812.00279 [cs.LG]
39. Aziguli ZY, Xie Y, Xu Y, Chen Y. 2017. Structural technology research on symptom data of Chinese medicine. In Proceedings of the IEEE 19th International Conference on e-Health Networking, Applications and Services. New York: IEEE
40. Shang J, Xiao C, Ma T, Li H, Sun J. 2019. GAMENet: graph augmented memory networks for recommending medication combination. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, pp. 1126–33. Palo Alto, CA: Assoc. Adv. Artif. Intell.
41. Huang EW, Wang S, Zhai C. 2017. VisAGE: integrating external knowledge into electronic medical record visualization. Proc. Pac. Symp. Biocomput 23:578–89
42. Singh A, Rawlings CJ, Hassani-Pak K. 2018. KnetMaps: a BioJS component to visualize biological knowledge networks. F1000Research 7:1651 [PubMed: 30755790]
43. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13(11):2498–504 [PubMed: 14597658]
44. Queralt-Rosinach N, Stupp GS, Li TS, Mayers M, Hoatlin ME, et al. 2019. Structured reviews for data and knowledge driven research. bioRxiv 729475. 10.1101/729475
45. Tripodi IJ, Callahan TJ, Westfall JT, Meitzer NS, Dowell RD, Hunter LE. 2019. Applying knowledge-driven mechanistic inference to toxicogenomics. bioRxiv 782011. 10.1101/782011
46. Smaili FZ, Gao X, Hoehndorf R. 2019. OPA2Vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction. Bioinformatics 35(12):2133–40 [PubMed: 30407490]
47. Celebi R, Yasar E, Uyar H, Gumus O, Dikenelli O, Dumontier M. 2018. Evaluation of knowledge graph embedding approaches for drug-drug interaction prediction using Linked Open Data. In Proceedings of the 2018 International Conference on Semantic Web Applications and Tools for Healthcare and Life Sciences, ed. Baker CJO, Waagmeester A, Splendiani A, Beyan AD, Marshall MS. CEUR Workshop Proc.
48. Crichton G, Guo Y, Pyysalo S, Korhonen A. 2018. Neural networks for link prediction in realistic biomedical graphs: a multi-dimensional evaluation of graph embedding-based approaches. BMC Bioinform. 19(1):176
49. Hamilton W, Bajaj P, Zitnik M, Jurafsky D, Leskovec J. 2018. Embedding logical queries on knowledge graphs. In Advances in Neural Information Processing Systems 31, ed. Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, pp. 2026–37. Red Hook, NY: Curran Assoc.

50. Jiang J, Wang H, Xie J, Guo X, Guan Y, Yu Q. 2018. Medical knowledge embedding based on recursive neural network for multi-disease diagnosis. arXiv:1809.08422 [cs.AI]
51. Karim MR, Cochez M, Jares JB, Uddin M, Beyan O, Decker S. 2019. Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-LSTM network. In Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, pp. 113–23. New York: Assoc. Comput. Mach.
52. Mohamed SK, Nounu A, Novaicek V. 2019. Drug target discovery using knowledge graph embeddings. In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, pp. 11–18. New York: Assoc. Comput. Mach.
53. Sadeghi A, Lehmann J. 2019. Linking physicians to medical research results via knowledge graph embeddings and Twitter. arXiv:1908.02571 [cs.SI]
54. Womack F, McClelland J, Koslicki D. 2019. Leveraging distributed biomedical knowledge sources to discover novel uses for known drugs. bioRxiv 765305. 10.1101/765305
55. Sang S, Yang Z, Liu X, Wang L, Lin H, et al. 2019. GrEDeL: a knowledge graph embedding based method for drug discovery from biomedical literatures. IEEE Access 7:8404–15
56. Tripodi I, Cohen KB, Hunter LE. 2017. A semantic knowledge-base approach to drug-drug interaction discovery. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, ed. Hu X, Gong Y, Shyu C-R, Korkin D, Bromberg Y, et al., pp. 1123–26. New York: IEEE
57. Li L, Wang P, Wang Y, Jiang J, Tang B, et al. 2019. PrTransH: embedding probabilistic medical knowledge from real world EMR data. arXiv:1909.00672 [cs.AI]
58. Yue X, Wang Z, Huang J, Parthasarathy S, Moosavinasab S, et al. 2019. Graph embedding on biomedical networks: methods, applications, and evaluations. arXiv:1906.05017 [cs.LG]
59. Callahan TJ. 2019. PheKnowLator. 10.5281/zenodo.3401437
60. Deng Y, Li Y, Shen Y, Du N, Fan W, et al. 2018. MedTruth: a semi-supervised approach to discovering knowledge condition information from multi-source medical data. arXiv:1809.10404 [cs.DB]
61. Morton K, Wang P, Bizon C, Cox S, Balhoff J, et al. 2019. ROBOKOP: an abstraction layer and user interface for knowledge graphs to support question answering. Bioinformatics 2019:btz604
62. Wright D, Katsis Y, Mehta R, Hsu C-N. 2019. NormCo: deep disease normalization for biomedical knowledge base construction. Paper presented at Automated Knowledge Base Construction Conference (AKBC 2019), Amherst, MA, May 20
63. Luan Y, He L, Ostendorf M, Hajishirzi H. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. arXiv:1808.09602 [cs.CL]
64. Zhang N, Deng S, Sun Z, Wang G, Chen X, et al. 2019. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks. arXiv:1903.01306 [cs.IR]
65. Duque A, Stevenson M, Martinez-Romo J, Araujo L. 2018. Co-occurrence graphs for word sense disambiguation in the biomedical domain. Artif. Intell. Med 87:9–19 [PubMed: 29573845]
66. Jin Z, Zhang Y, Kuang H, Yao L, Zhang W, Pan Y. 2019. Named entity recognition in traditional Chinese medicine clinical cases combining BiLSTM-CRF with knowledge graph. In Knowledge Science, Engineering, and Management, ed. Douligeris C, Karagiannis D, Apostolou D, pp. 537–48. Cham, Switz.: Springer
67. Logan R, Liu NF, Peters ME, Gardner M, Singh S. 2019. Barack's wife Hillary: using knowledge graphs for fact-aware language modeling. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ed. Korhonen A, Traum D, Màrquez L, pp. 5962–71. New York: Assoc. Comput. Linguist.
68. Wang Z, Xu S, Zhu L. 2018. Semantic relation extraction aware of N-gram features from unstructured biomedical text. J. Biomed. Inform 86:59–70 [PubMed: 30172761]
69. Xie Y, Yan C, Zhang D. 2018. Personalized diagnostic modal discovery of traditional Chinese medicine knowledge graph. In Proceedings of the 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, pp. 1096–103. New York: IEEE
70. Gao Z, Fu G, Ouyang C, Tsutsui S, Liu X, et al. 2019. edge2vec: representation learning using edge semantics for biomedical knowledge discovery. BMC Bioinform. 20(1):306

71. Su C, Tong J, Zhu Y, Cui P, Wang F. 2018. Network embedding in biomedical data science. Brief. Bioinform 2018:bby117
72. Sang S, Yang Z, Wang L, Liu X, Lin H, Wang J. 2018. SemaTyP: a knowledge graph based literature mining method for drug discovery. BMC Bioinform. 19(1):193
73. Vlietstra WJ, Vos R, Sijbers AM, van Mulligen EM, Kors JA. 2018. Using predicate and provenance information from a knowledge graph for drug efficacy screening. J. Biomed. Semant 9(1):23
74. Wang Q, Wang T, Xu C. 2018. Using a knowledge graph for hypernymy detection between Chinese symptoms. In Proceedings of the Tenth International Conference on Advanced Computational Intelligence, pp. 601–6. New York: IEEE
75. Zhang D, He D, Zou N, Zhou X, Pei F. 2018. Automatic relationship verification in online medical knowledge base: a large scale study in SemMedDB. In Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine, pp. 1673–80. New York: IEEE
76. Perez N, Cuadros M, Rigau G. 2018. Biomedical term normalization of EHRs with UMLS. arXiv:1802.02870 [cs.CL]
77. Sharma S, Santra B, Jana A, Santosh TYSS, Ganguly N, Goyal P. 2019. Incorporating domain knowledge into medical NLI using knowledge graphs. arXiv:1909.00160v1 [cs.CL]
78. Wang X, Li Q, Ding X, Zhang G, Weng L, Ding M. 2019. A new method for complex triplet extraction of biomedical texts. In Knowledge Science, Engineering, and Management, ed. Douligeris C, Karagiannis D, Apostolou D, pp. 146–58. Cham, Switz.: Springer
79. Huang L, Yu C, Chi Y, Qi X, Xu H. 2019. Towards smart healthcare management based on knowledge graph technology. In Proceedings of the 2019 8th International Conference on Software and Computer Applications, pp. 330–37. New York: Assoc. Comput. Mach.
80. Fauqueur J, Thillaisundaram A, Togia T. 2019. Constructing large scale biomedical knowledge bases from scratch with rapid annotation of interpretable patterns. arXiv:1907.01417 [cs.CL]
81. Cong Q, Feng Z, Li F, Zhang L, Rao G, Tao C. 2018. Constructing biomedical knowledge graph based on SemMedDB and linked open data. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, ed. Zheng H, Hu X, Callejas Z, Schmidt H, Griol D, et al., pp. 1628–31. New York: IEEE
82. Brandizi M, Singh A, Rawlings C, Hassani-Pak K. 2018. Towards FAIRer biological knowledge networks using a hybrid linked data and graph database approach. J. Integr. Bioinform 15(3):20180023
83. Biomedical Data Transl. Consort. 2019. Toward a universal biomedical data translator. Clin. Transl. Sci 12(2):86–90 [PubMed: 30412337]
84. Al-Nagdawi A 2019. Building the ultimate nexus of knowledge for biomedical data. Web Resource, Univ. Calif. San Franc. https://precisionmedicine.ucsf.edu/building-ultimate-nexus-knowledge-biomedical-data
85. Nelson CA, Butte AJ, Baranzini SE. 2019. Integrating biomedical research and electronic health records to create knowledge-based biologically meaningful machine-readable embeddings. Nat. Commun 10(1):3045 [PubMed: 31292438]
86. UCSF (Univ. Calif. San Franc.). 2019. What is SPOKE? Web Resource, Univ. Calif. San Franc. https://spoke.ucsf.edu/
87. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, et al. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife 6:e26726 [PubMed: 28936969]
88. Yuan J, Jin Z, Guo H, Jin H, Zhang X, et al. 2019. Constructing biomedical domain-specific knowledge graph with minimum supervision. Knowl. Inform. Syst 10.1007/s10115-019-01351-4
89. NCATS (Natl. Cent. Adv. Transl. Sci.). 2019. Biomedical Data Translator: development. Notice Funding Oppor. NOT-TR-19-028, Natl. Cent. Adv. Transl. Sci., Bethesda, MD. https://ncats.nih.gov/files/NCATS-Translator-FY20-COMBINED-FOA-FINAL.pdf
90. Google. 2019. Welcome to Data Commons. http://datacommons.org
91. Belleau F, Nolin M-A, Tourigny N, Rigault P, Morissette J. 2008. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J. Biomed. Inform 41(5):706–16 [PubMed: 18472304]

92. Nolin M-A, Ansell P, Belleau F, Idehen K, Rigault P, et al. 2008. Bio2RDF network of linked data. Paper presented at Semantic Web Challenge: International Semantic Web Conference (ISWC 2008), Karlsruhe, Ger., Oct. 26–30

93. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, et al. 2017. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 45(D1):D833–39 [PubMed: 27924018]

94. Messina A, Pribadi H, Stichbury J, Bucci M, Klarman S, Urso A. 2018. BioGrakn: a knowledge graph-based semantic database for biomedical sciences. In Complex, Intelligent, and Software Intensive Systems, ed. Barolli L, Terzo O, pp. 299–309. Cham, Switz.: Springer

95. Lossio-Ventura JA, Hogan W, Modave F, Guo Y, He Z, et al. 2018. OC-2-KB: integrating crowdsourcing into an obesity and cancer knowledge base curation system. BMC Med. Inform. Decis. Making 18(Suppl. 2):55

96. Ali M, Hoyt CT, Domingo-Fernández D, Lehmann J, Jabeen H. 2019. BioKEEN: a library for learning and evaluating biological knowledge graph embeddings. Bioinformatics 35(18):3538–40 [PubMed: 30768158]

97. Musen MA, Protégé Team. 2015. The Protégé Project: a look back and a look forward. AI Matters 1(4):4–12 [PubMed: 27239556]

98. Systap LLC. 2013. The bigdata® RDF database. White Pap., BlazeGraph Database, Washington, DC. https://blazegraph.com/docs/bigdata_architecture_whitepaper.pdf

99. Smith B, Ashburner M, Rosse C, Bard J, Bug W, et al. 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol 25(11):1251–55 [PubMed: 17989687]

100. RDF Data Access Work. Group. 2008. SPARQL query language for RDF. Web Resour., World Wide Web Consort. https://www.w3.org/TR/rdf-sparql-query/

101. Page RDM. 2019. Ozymandias: a biodiversity knowledge graph. PeerJ 7:e6739 [PubMed: 30993051]

102. Page R. 2018. Liberating links between datasets using lightweight data publishing: an example using plant names and the taxonomic literature. Biodivers. Data J 6:e27539

103. Senderov V, Simov K, Franz N, Stoev P, Catapano T, et al. 2018. OpenBiodiv-O: ontology of the Open-Biodiv knowledge management system. J. Biomed. Semant 9(1):5

104. Badal VD, Wright D, Katsis Y, Kim H-C, Swafford AD, et al. 2019. Challenges in the construction of knowledge bases for human microbiome-disease associations. Microbiome 7(1):129 [PubMed: 31488215]

105. Ammar W, Groeneveld D, Bhagavatula C, Beltagy I, Crawford M, et al. 2018. Construction of the literature graph in Semantic Scholar. arXiv:1805.02262 [cs.CL]

106. Auer S, Kovtun V, Prinz M, Kasprzik A, Stocker M, Vidal ME. 2018. Towards a knowledge graph for science. In Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, Art. 1. New York: Assoc. Comput. Mach.

107. Dai Q, Inoue N, Reisert P, Takahashi R, Inui K. 2019. Distantly supervised biomedical knowledge acquisition via knowledge graph based attention. In Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications, pp. 1–10. New York: Assoc. Comput. Mach.

108. Jiang T, Zhao T, Qin B, Liu T, Chawla NV, Jiang M. 2019. The role of "condition": a novel scientific knowledge graph representation and construction model. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1634–42. New York: Assoc. Comput. Mach.

109. Buscaldi D, Dessì D, Motta E, Osborne F, Reforgiato Recupero D. 2019. Mining scholarly publications for scientific knowledge graph construction. In Proceedings of the Extended Semantic Web Conference, ed. Hitzler P, Kirrane S, Hartig O, de Boer V, Vidal M-E, pp. 8–12. Cham, Switz.: Springer

110. Tennakoon C, Zaki N, Arnaout H, Elbassuoni S, El-Hajj W, Al Jaberi A. 2019. Biological knowledge graph construction, search, and navigation. In Leveraging Biomedical and Healthcare Data, ed. Kobeissy F, Alawieh A, Zaraket FA, Wang K, pp. 107–20. London: Academic

111. Cheng B, Zhang Y, Cai D, Qiu W, Shi D. 2018. Construction of traditional Chinese medicine knowledge graph using data mining and expert knowledge. In IEEE International Conference on Network Infrastructure and Digital Content, pp. 209–13. New York: IEEE

112. Gong F, Chen Y, Wang H, Lu H. 2019. On building a diabetes centric knowledge base via mining the web. BMC Med. Inform. Decis. Making 19(Suppl. 2):49

113. Gyrard A, Gaur M, Shekarpour S, Thirunarayan K, Sheth A. 2018. Personalized health knowledge graph. In Proceedings of the 1st Workshop on Contextualized Knowledge Graphs. http://knoesis.wright.edu/sites/default/files/personalized-asthma-obesity%20%2814%29.pdf

114. Wang H, Miao X, Yang P. 2018. Design and implementation of personal health record systems based on knowledge graph. In Proceedings of the 9th International Conference on Information Technology in Medicine and Education, pp. 133–36. Los Alamitos, CA: IEEE Comput. Soc.

115. Gao M, Wang R, Suny D, Li G. 2018. Intelligent healthcare knowledge resources in Chinese: a survey. In Proceedings of the 15th International Symposium on Pervasive Systems, Algorithms and Networks, pp. 318–24. Los Alamitos, CA: IEEE Comput. Soc.

116. Li D. 2018. Modelling online user behavior for medical knowledge learning. Ind. Manag. Data Syst 118(4):889–911

117. Nordon G, Koren G, Shalev V, Kimelfeld B, Shalit U, Radinsky K. 2019. Building causal graphs from medical literature and electronic medical records. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, pp. 1102–9. Palo Alto, CA: Assoc. Adv. Artif. Intell.

118. Xia E, Sun W, Mei J, Xu E, Wang K, Qin Y. 2018. Mining disease-symptom relation from massive biomedical literature and its application in severe disease diagnosis. AMIA Annu. Symp. Proc 2018:1118–26 [PubMed: 30815154]

119. Fang Y, Wang H, Wang L, Di R, Song Y. 2019. Diagnosis of COPD based on a knowledge graph and integrated model. IEEE Access 7:46004–13

120. Zhang H, He X, Harrison T, Bian J. 2019. Aero: an evidence-based semantic web knowledge base of cancer behavioral risk factors. In Proceedings of the 4th International Workshop on Semantics-Powered Data Mining and Analytics, ed. He Z, Bian J, Tao C, Zhang R, pp. 7–11. CEUR Workshop Proc.

121. He X, Zhang R, Rizvi R, Vasilakes J, Yang X, et al. 2019. ALOHA: developing an interactive graph-based visualization for dietary supplement knowledge graph through user-centered design. BMC Med. Inform. Decis. Making 19(Suppl. 4):150

122. Nordon G, Koren G, Shalev V, Horvitz E, Radinsky K. 2019. Separating wheat from chaff: joining biomedical knowledge and patient data for repurposing medications. In Proceedings of the 33rd AIAA Conference on Artificial Intelligence, pp. 9565–72. Palo Alto, CA: Assoc. Adv. Artif. Intell.

123. Shen Y, Yuan K, Dai J, Tang B, Yang M, Lei K. 2019. KGDDS: a system for drug-drug similarity measure in therapeutic substitution based on knowledge graph curation. J. Med. Syst 43(4):92 [PubMed: 30834481]

124. Agibetov A, Jiménez-Ruiz E, Ondrésik M, Solimando A, Banerjee I, et al. 2018. Supporting shared hypothesis testing in the biomedical domain. J. Biomed. Semantics 9(1):9 [PubMed: 29422110]

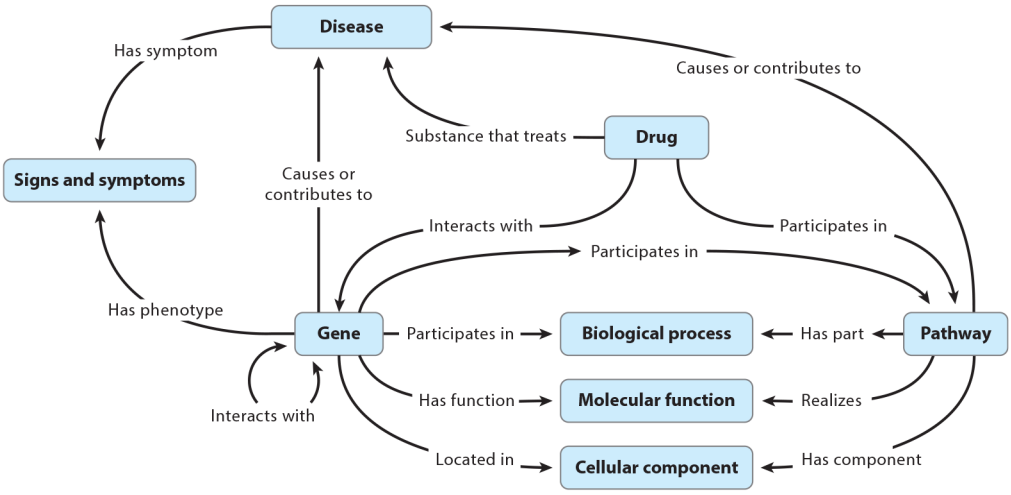

Figure 1.

An example of a knowledge representation for building a biomedical knowledge graph. Boxes represent different types of data, which are drawn from ontologies and other sources of linked open data. Boxes are connected by directed edges and represent semantically and biologically meaningful relationships.

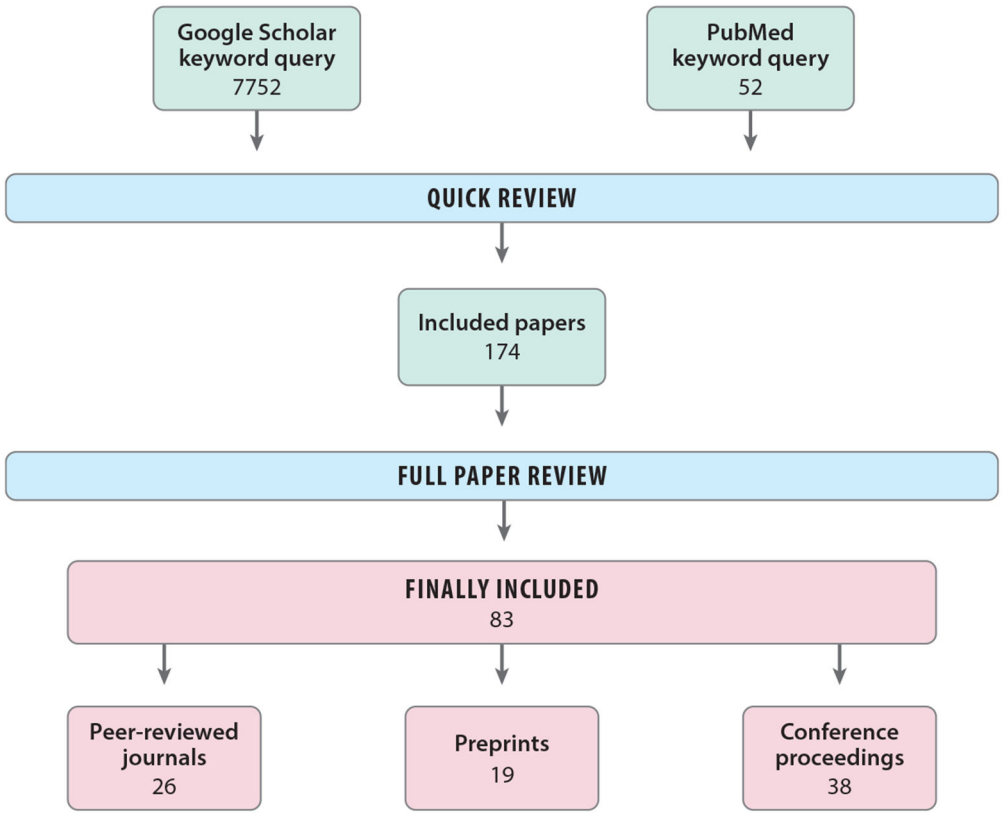

Figure 2.

Paper selection process outline. Combining the results from PubMed and Google Scholar queries, we narrowed down the list of papers using a two-step process. First, we performed a quick review to reduce the initial number of papers. Then, we closely inspected each paper, which helped us to arrive at the final set of 83 papers.

Table 1

Currently available biomedical knowledge graphs

Name (Reference)	Primitives[a]	Domain	Backend	Last updated	Construction method[b]
Bio2RDF (91)	Mixed concepts	Biomedical	Virtuoso	09/25/14	Mixed integration
BioGrakn (95)	Mixed concepts	Biomedical	Grakn.ai	09/27/19	Identifier-based integration
Data Commons Graph (90)	Concepts	Demographics, health, economics, crime, education, employment, housing	GO, Python	2019	Identifier-based integration
DisGeNET (94)	Ontology concepts (URIs)	Biomedical	RDF	01/01/19	Mixed integration
Hetionet (87)	Mixed concepts	Biomedical	Neo4j	07/08/19	Mixed integration
KaBOB (37)	Ontology concepts (URIs)	Biomedical	AllegroGraph	06/23/19	Semantic integration of OBOs
NGLY1 deficiency (44)	Mixed concepts	NGLY1 deficiency	Neo4j	08/08/19	Mixed integration
Ozymandias (101)	Ontology concepts (URIs)	Biodiversity	Blazegraph	2019	Linked data from ALA and ALD using CrossRef
PheKnowLator (59)	Mixed concepts	Biomedical	RDF, Python	09/25/19	Mixed integration
ROBOKOP (61)	Mixed concepts	Biomedical	GreenT, Neo4j	09/20/19	Mixed integration
SemMedDB (19)	UMLS concepts	Biomedical	Relational database	06/2018	Mixed integration; Semantic Knowledge Representation program and the UMLS
Sparklis (32)	Ontology concepts (URIs)	Pharmacovigilance	Neo4j, JSON API	01/2019	Semantic integration of OBOs
SPOKE (86)	Mixed concepts	Biomedical	Neo4j	2019	Mixed integration; extends Hetionet with other data in UCSF Information Commons

[a] Mixed concepts are constructed from ontology concepts and other nonontology concepts found in sources of biomedical data.

[b] Created using semantic integration (i.e., ontology-based) methods and methods that integrate data by matching up sets of identifiers (i.e., using other data sources like linked data, with standardized concept identifiers).

Abbreviations: ALA, Atlas of Living Australia; ALD, Australian Faunal Directory; API, application programming interface; GO, Gene Ontology; JSON, JavaScript Object Notation; NGLY1, *N*-glycanase 1; OBO, Open Biomedical and Biological Ontology; RDF, Resource Description Framework; UMLS, Unified Medical Language System; URI, uniform resource identifier.