Title

Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science

Permalink

https://escholarship.org/uc/item/5zf5q22w

Journal

Clinical and Translational Science, 15(8)

ISSN

1752-8054

Authors

Unni, Deepak R
Moxon, Sierra AT
Bada, Michael
et al.

Publication Date

2022-08-01

DOI

10.1111/cts.13302

Peer reviewed

Received: 4 March 2022 | Revised: 27 April 2022 | Accepted: 2 May 2022

DOI: 10.1111/cts.13302

MINI-REVIEW

Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science

Deepak R. Unni[1,2] | Sierra A. T. Moxon[2] | Michael Bada[3] | Matthew Brush[3] | Richard Bruskiewich[4] | J. Harry Caufield[2] | Paul A. Clemons[5] | Vlado Dancik[5] | Michel Dumontier[6] | Karamarie Fecho[7] | Gustavo Glusman[8] | Jennifer J. Hadlock[8] | Nomi L. Harris[2] | Arpita Joshi[8] | Tim Putman[3] | Guangrong Qin[8] | Stephen A. Ramsey[9] | Kent A. Shefchek[3] | Harold Solbrig[10] | Karthik Soman[11] | Anne E. Thessen[3] | Melissa A. Haendel[3] | Chris Bizon[7] | Christopher J. Mungall[2] | The Biomedical Data Translator Consortium[†]

[1]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany

[2]Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, California, USA

[3]Center for Health AI, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA

[4]Star Informatics, Sooke, British Columbia, Canada

[5]Chemical Biology and Therapeutics Science Program, Broad Institute, Cambridge, Massachusetts, USA

[6]Institute of Data Science, Maastricht University, Maastricht, The Netherlands

[7]Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

[8]Institute for Systems Biology, Seattle, Washington, USA

[9]Department of Biomedical Sciences, Oregon State University, Corvallis, Oregon, USA

[10]Johns Hopkins University, Baltimore, Maryland, USA

[11]Department of Neurology, University of California San Francisco, San Francisco, California, USA

Correspondence
Christopher J. Mungall, Lawrence Berkeley National Laboratory, One Cyclotron Road, MS 977, Berkeley CA 94720, USA.
Email: cjmungall@lbl.gov

Funding information
This work was supported in part by the NCATS Biomedical Data Translator program (Other Transaction Awards OT2TR003434, OT2TR003436, OT2TR003428, OT2TR003448, OT2TR003427, OT2TR003430,

Abstract

Within clinical, biomedical, and translational science, an increasing number of projects are adopting graphs for knowledge representation. Graph-based data models elucidate the interconnectedness among core biomedical concepts, enable data structures to be easily updated, and support intuitive queries, visualizations, and inference algorithms. However, knowledge discovery across these "knowledge graphs" (KGs) has remained difficult. Data set heterogeneity and complexity; the proliferation of ad hoc data formats; poor compliance with guidelines on findability, accessibility, interoperability, and reusability; and, in particular, the lack of a universally accepted, open-access model for standardization

Deepak R. Unni and Sierra A. T. Moxon served as co-lead authors.

[†]Consortial/collaborative authors.

OT2TR003433, OT2TR003450, OT2TR003437, OT2TR003443, OT2TR003441, OT2TR003449, OT2TR003445, OT2TR003422, OT2TR003435, OT3TR002026, OT3TR002020, OT3TR002025, OT3TR002019, OT3TR002027, OT2TR002517, OT2TR002514, OT2TR002515, OT2TR002584, and OT2TR002520; Contract number 75N95021P00636). Additional funding was provided by the Office of the Director, National Institutes of Health (grant award R24-OD011883), the National Human Genome Research Institute (grant award 7RM1HG010860-02), and the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC0205CH11231.

across biomedical KGs has left the task of reconciling data sources to downstream consumers. Biolink Model is an open-source data model that can be used to formalize the relationships between data structures in translational science. It incorporates object-oriented classification and graph-oriented features. The core of the model is a set of hierarchical, interconnected classes (or categories) and relationships between them (or predicates) representing biomedical entities such as gene, disease, chemical, anatomic structure, and phenotype. The model provides class and edge attributes and associations that guide how entities should relate to one another. Here, we highlight the need for a standardized data model for KGs, describe Biolink Model, and compare it with other models. We demonstrate the utility of Biolink Model in various initiatives, including the Biomedical Data Translator Consortium and the Monarch Initiative, and show how it has supported easier integration and interoperability of biomedical KGs, bringing together knowledge from multiple sources and helping to realize the goals of translational science.

INTRODUCTION

The use of graphs to formalize the representation of human knowledge dates back to the origins of artificial intelligence and the use of semantic networks for knowledge representation.[1,2] The term "knowledge graph" (KG) is gaining popularity and is generally used to encompass a range of graph-oriented representation frameworks, including Resource Description Framework (RDF) triple stores and labeled property-graph databases, such as Neo4j. Examples of general-domain KGs include the Google Knowledge Graph and Wikidata.[3] Within the biomedical sciences, examples include SemMedDB,[4] Hetionet,[5] Implicitome,[6] Monarch Initiative,[7] the biological subset of Wikidata,[8] SPOKE,[9] and KG-COVID-19.[10]

Although KGs have been defined in various ways, perhaps the most intuitive definition is a graph in which the nodes represent real-world entities and the edges represent known relationships between those entities.[11] In a KG, the knowledge or "facts" are represented as statements, with each statement modeled as two nodes linked together by an edge representing the relationship between them. The statements can have additional properties, metadata, and qualifying attributes that further capture the meaning of the statement and characterize the properties of nodes and edges.

Because the basic structure of a KG is generic, the knowledge contained within a KG can be heterogeneous and mutable and still be representable in the graph. The representation of knowledge as simple connections between core entities makes iterative, rapid development of KGs possible. In addition, by leveraging the graph data structure and using various inference strategies, one can infer new edges or connections between nodes in a graph. Ontology-oriented KGs allow deductive inference through logical rules, from basic rules such as the Gene Ontology "true path" rule[12] to more sophisticated methods like Description Logic inference.[13] Ontology-oriented KGs are also amenable to machine learning approaches, such as embedding in vector space,[14] which supports the application of deep neural networks for tasks such as link prediction and node classification. Within the biomedical sciences, ontology-oriented KGs have been used for tasks, such as drug repurposing,[5] target prioritization,[15] and phenotype profile matching.[7]

Several ontologies and schemas for representing biomedical knowledge are available. A constellation of domain-specific ontologies from the Open Biological and biomedical Ontology Foundry[16] can be used for modeling knowledge. For example, the Semantic Science Integrated Ontology[17] is used for representing scientific data and knowledge. The Wikidata Ontology[18] is used by Wikidata for representing knowledge. In terms of schemas, schema.org is used for representing metadata about entities and relationships to other entities. BioSchemas is an extension of schema.org for representing metadata about biological entities.

Whereas existing efforts in modeling knowledge have been valuable, a unified data model that bridges across multiple ontologies, schemas, and data models does not exist. Here, we present Biolink Model as an open-source, universal data model that defines entities and the relationships between these entities within translational science.

OVERVIEW OF BIOLINK MODEL

Biolink Model is a data model for organizing data in biomedical KGs. The model serves both as a map for bringing together data from different sources under one unified model, and as a bridge between ontological domains.

Biolink Model is composed of several modeling elements, including a hierarchy of defined classes, properties (with defined types), predicates, mixins, and associations (Table 1). Domain knowledge in a KG that conforms to Biolink Model is represented using associations. An association minimally includes a subject and an object (Biolink Model classes) related by a Biolink Model predicate, together comprising its core triple (statement or primary assertion). The subject and object of an association are foundational domain concepts (e.g., genes, diseases, chemicals, and phenotypes), whose Internationalized Resource Identifiers (IRIs) come from community standard ontologies (e.g., HGNC, MONDO, ChEBI, and HPO). The predicate is a Biolink Model element that represents the relationship between the subject and object. Associations may also include slots to hold additional metadata about the core triple, primarily information about the provenance, and evidence supporting the assertion (Figure 1).

Biolink Model aims to address several challenges that obstruct the interoperability between KGs, including: (1) the need for expertise to transform data between tabular, RDF, and graphical models; (2) sparse and/or inconsistent application of ontologies or other controlled vocabularies, as well as differences in the identifiers that are used for storing instances of nodes within a graph; and (3) the lack of a standard approach to model the intersection of ontological domains (e.g., the relationships between genes and diseases).

Using the framework provided by the Linked data Modeling Language (LinkML), Biolink Model is distributed in a variety of formats, including YAML, JSON-Schema,

TABLE 1 Biolink Model elements and their definitions

Biolink Model element	Definition	Examples
Class	High-level types (or categories) representing core biological concepts of interest such as genes, diseases, chemical substances, anatomic structures, and phenotypic features, arranged in a class hierarchy	biolink:Disease, biolink:PhenotypicFeature, biolink:Gene, biolink:SequenceVariant
Predicate	Objects that define the action being carried out by the subject (or named entity) of a core triple and help define how two entities (or classes) can be related to one another. In graph formalism, predicates are relationships that link two instances. Predicates in the Biolink Model all descend from the "biolink:related_to" predicate	biolink:has_phenotype, biolink:positively_regulates, biolink:affects, biolink:associated_with, biolink:related_to
Node property	A set of attributes that can be regarded as a characteristic or inherent part of an instance of "biolink:NamedThing"	biolink:symbol, biolink:name, biolink:id
Edge property	A set of attributes that can be regarded as a characteristic or inherent part of a statement, association, or edge	biolink:publications, biolink:has_evidence
Core triple	The domain knowledge of an association expressed by the subject and object nodes plus the predicate connecting them	biolink:Disease biolink:has_phenotype biolink:PhenotypicFeature
Association	Associations are classes that define a relationship between two domain concepts, constrained and qualified by edge attributes	biolink:DiseaseToPhenotypicFeatureAssociation, biolink:GeneToDiseaseAssociation
Type	A kind of value that tells what operations can be performed on a particular data set. Biolink Model implements common types, such as integer and string, but it also defines custom types like quotient and unit	URI or CURIE, string, integer, biolink:Quotient, biolink:Unit
Mixin	Modeling elements used to extend the properties (or slots) of a class, without changing its position in the class hierarchy. Please see the Biolink Model documentation for more information on mixin elements	biolink:GeneOrGeneProduct, biolink:DiseaseOrPhenotypicFeature

Abbreviations: CURIE, compact URI; URI, unique resource identifier.

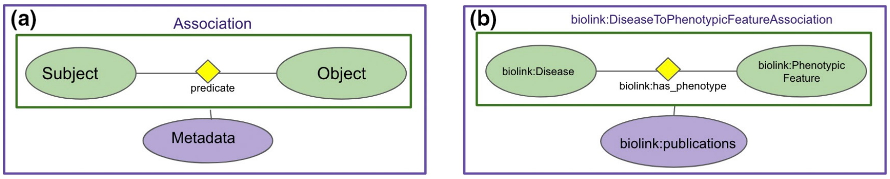

FIGURE 1 An example of an Association represented in Biolink Model. In (a), the green ovals represent the subject and object classes, connected by a predicate. Together, the classes and the predicate constitute a statement or "core triple" in the model. Edge properties provide further context and qualification to the core triple. The entire diagram, including the core triple and its provenance, represents a Biolink Model "association." In (b), we see a specific example of a "biolink:DiseaseToPhenotypicFeatureAssociation," where the subject is "biolink:Disease," the object is "biolink:PhenotypicFeature," and the predicate is "biolink:has_phenotype." In addition, the "biolink:publications" property (lavender oval) records the provenance of the core triple.

SQL-DDL, Python/Java classes, and RDF. Additionally, Unified Modeling Language diagrams provide a visual representation of the model. Biolink Model is accessible in frameworks familiar to a wide variety of developers and database engineers. Because the model can be distributed in different formats, the model elements can also be validated using existing toolchains (e.g., JSONSchema validation and SQL constraints), thus speeding up the reconciliation of tabular data, ontologies, and graphs.

The biomedical field has been a leader and champion of ontology development. However, this has sometimes led to the development of multiple ontologies or controlled vocabularies for the same domain concept. When this happens, KG creators must identify which vocabulary best suits their needs, as well as understand how to apply concepts from the chosen ontology to their class instances. Biolink Model helps solve this challenge by indicating to users which ontologies should be used for instances of its classes via identifier prefixes (id_prefixes), mappings, and associations.

Biolink Model describes its classes in a description field. Part of the definition of a class is an id_prefixes construct. Recognizing that biomedical resources often implement new identifiers for their resource, instead of reusing existing identifiers from other resources,[19] Biolink Model encourages reuse of existing ontologies by providing a list of possible ontologies (via id_prefixes) in preference order for engineers to use when instantiating model classes. For example, for a disease class, Biolink Model suggests that instances of the class use Mondo (the Mondo Disease Ontology)[20] as the preferred disease vocabulary. The id_prefixes modeling construct allows the development of software that can normalize identifiers across data sources. Tools such as the Biomedical Data Translator Node Normalization Service and the Knowledge Graph Exchange Framework use the identifier mappings in Biolink Model to return the preferred equivalent identifier when presented with several identifiers that represent the same domain concept but with different namespaces (e.g., NCBIGene vs. HGNC gene identifiers).

Each element in Biolink Model is mapped, when possible, to equivalent elements in other ontologies or models. Biolink Model uses mapping terms from the Simple Knowledge Organization System (SKOS) namespace to record classes and objects outside the model that can be considered similar in an exact, broad, narrow, close, or related manner to the Biolink Model class (e.g., the broad_mapping relation implements the skos:broadMatch). These mappings render the model and data more computable, allowing software programs to automatically harmonize and connect disparate data sources, thus facilitating interoperability.

Finally, a key feature of Biolink Model is its association elements. Taking inspiration from successful efforts like Semanticscience Integrated Ontology,[17] Biolink Model Association elements establish rules for transforming biomedical knowledge into computable statements and help define how to represent knowledge statements across ontological domains. "Computable," in this context, means that each Biolink Model association defines the kinds of objects that can participate as a subject or object of a biomedical statement (via domain and range constraints); defines sets of attributes (edge properties described in Table 1 and detailed in the Biolink Model documentation) that are required to properly instantiate a relationship between two domain concepts; and provides a blueprint for registering and maintaining the provenance of each statement. In Web Ontology Language (OWL),[21] Biolink Model association elements are equivalent to axioms, and in RDF, they are equivalent to statements (rdf:Statement). Because provenance and evidence are critical components of any data set (and the knowledge represented therein), Biolink Model provides properties capable of tracking evidence and provenance both at the class and association levels.

APPLICATIONS OF BIOLINK MODEL

Translational science, by its nature, involves the application of diverse information derived from different subject matter experts and curated data sources to answer questions through integrated analyses of clinical and biomedical knowledge. Biolink Model supports translation, integration, and harmonization across knowledge sources by capturing subject matter expertise in a machine-readable format that allows software to interoperate with disparate data sources using a common dialect, facilitated by a harmonized data model.

We highlight several examples here.

Biomedical Data Translator ("Translator") Consortium

The Translator Consortium[22] has adopted Biolink Model as an open-source upper-level data model that supports semantic harmonization and reasoning across diverse Translator "knowledge sources."[15] The model serves a central role in the Translator program and forms the architectural basis of the Translator system, as described below.

The Translator program aims to develop a comprehensive, relational, N-dimensional infrastructure designed to integrate disparate data sources—including objective signs and symptoms of disease, drug effects, chemical and genetic interactions, cell and organ pathology, and other relevant biological entities and relations—and reason over the integrated data to rapidly derive biomedical insights.[23] The ultimate goal of Translator is to augment human reasoning and thereby accelerate translational science and knowledge discovery.

To achieve its ambitious goal, the Translator project assembled a diverse interdisciplinary team and a variety of biomedical data sources, including electronic health record data, clinical trial data, genomic and other -omics data, chemical reaction data, and drug data. There are hundreds of data sources in the Translator ecosystem, each of which had its own data representation and were in formats that were not compatible or interoperable. Moreover, groups within the Translator Consortium had integrated the data sources as knowledge sources within independent KGs, but these KGs were developed using different technologies and formalisms, such as property graphs in Neo4j and semantically linked data via RDF and OWL.

In order to interoperate between knowledge sources and reason across KGs, Biolink Model was adopted as the common dialect, thus enabling queries over the entire Translator KG ecosystem. The result was a federated, harmonized ecosystem that supports advanced reasoning and inference to derive biomedical insights based on user queries.

An example Translator use case involved a collaboration with investigators at the Hugh Kaul Precision Medicine Institute (PMI) at the University of Alabama at Birmingham. PMI investigators posed the following natural-language question to the Translator Consortium: *what chemicals or drugs might be used to treat neurological disorders, such as epilepsy that are associated with genomic variants of RHOBTB2*? The investigators noted that *RHOBTB2* variants cause an accumulation of RHOBTB2

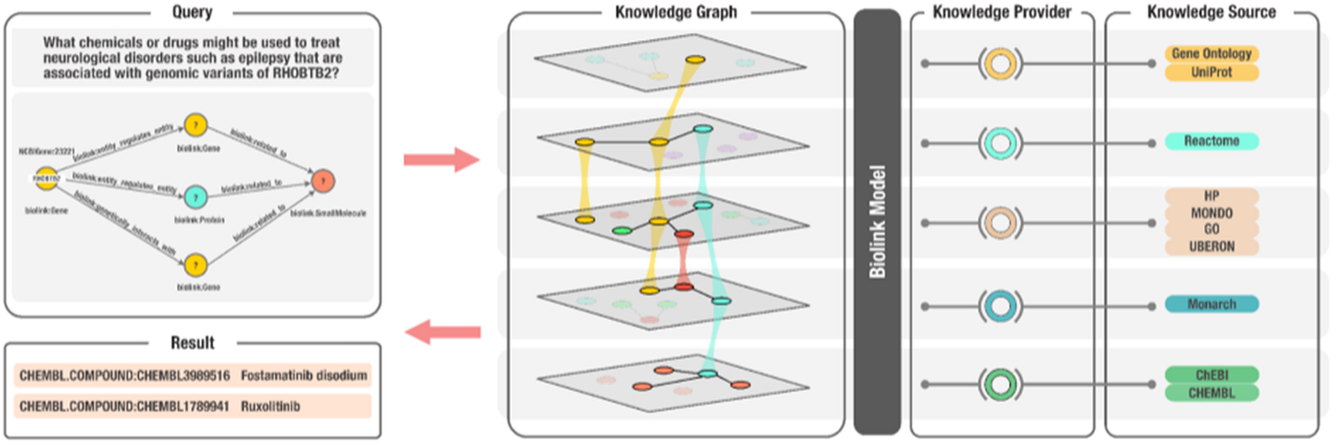
Query
What chemicals or drugs might be used to treat neurological disorders such as epilepsy that are associated with genomic variants of RHOBTB2?
NCBIGene:23221
RHOBTB2
biolink:Gene
biolink:entity_regulates_entity
biolink:entity_regulates_entity
biolink:genetically_interacts_with
?
biolink:Gene
?
biolink:Protein
?
biolink:Gene
biolink:related_to
biolink:related_to
biolink:related_to
?
biolink:SmallMolecule
Result
CHEMBL.COMPOUND:CHEMBL3989516 Fostamatinib disodium
CHEMBL.COMPOUND:CHEMBL1789941 Ruxolitinib
Knowledge Graph
Biolink Model
Knowledge Provider
Knowledge Source
Gene Ontology
UniProt
Reactome
HP
MONDO
GO
UBERON
Monarch
ChEBI
CHEMBL

FIGURE 2 An overview of the Translator architecture that supports biomedical KG-based question-answering, including the role of Biolink Model, in the context of an example question. In this example, a user has posed the natural-language question: what chemicals or drugs might be used to treat neurological disorders, such as epilepsy, that are associated with genomic variants of RHOBTB2? The question is translated into a graph query, as shown in the top left panel, which is then translated into a Translator standard machine query (not shown). The KG shown in the second panel from the left is derived from a variety of diverse "knowledge sources," a subset of which are displayed in the figure, that are exposed by Translator "knowledge providers." Biolink Model provides standardization and semantic harmonization across the disparate knowledge sources, thereby allowing them to be integrated into a KG capable of supporting question-answering. In this example, Translator provided two answers or results of interest to the investigative team who posed the question, namely, fostamatinib disodium and ruxolitinib, as shown in the bottom left panel. KG, knowledge graph.

protein and that this accumulation is believed to be the cause of the neurological disorder.

To answer the PMI investigator's question, Translator team members structured the following query: *NCBIGene:23221* (CURIE for RHOBTB2) -> *[biolink:entity_regulates_entity, biolink:genetically_interacts_with]* -> *biolink:Protein, biolink:Gene* -> *[biolink:related_to]* -> *biolink:SmallMolecule* (Figure 2). Because of the hierarchical structure of the Biolink model, the use of *biolink:related_to* also will return more specific predicates such as *biolink:negatively_regulates* and *biolink:positively_regulates*. The objective was to identify drugs or chemicals that might regulate *RHOBTB2* in some manner and thereby reduce the variant-induced accumulation of RHOBTB2 and associated neurological symptoms. As all nodes and edges within the Translator KG ecosystem are annotated to Biolink Model classes and attributes, a Translator query can be constructed from a natural-language user question and return results across a multitude of independent data sources. In addition, because the model uses hierarchical classes, with inheritance and polymorphism, natural-language queries translated to graph queries using Biolink Model syntax can be constructed at varying levels of granularity and return results from all levels of the hierarchy. Finally, because Biolink Model provides attributes on both edges and nodes that record provenance and evidence for these knowledge statements, each result is annotated with the trail of evidence that supports it.

When Translator team members sent the query to the Translator system, it returned several candidates of interest to PMI investigators, including fostamatinib disodium (CHEMBL.COMPOUND:CHEMBL3989516) and ruxolitinib (CHEMBL.COMPOUND:CHEMBL1789941). A review of the supporting evidence provided by Translator indicates that these are approved drugs that either directly or indirectly reduce or otherwise regulate the expression of *RHOBTB2*. Thus, Biolink Model helped Translator teams bring data together into a single system, thereby reducing the burden on the user to find and manually assemble data from these independent resources.

Monarch initiative

Similar to Translator, the Monarch Initiative is a large-scale bioinformatics web resource focused on leveraging existing biomedical knowledge to connect genotypes with phenotypes in an effort to aid research on genetic diseases. Monarch pulls together data from a wide variety of sources. However, because each source uses its own model to describe entities and their relationships, subject matter expertise is required to manually translate between knowledge representations. Monarch is adopting the Biolink Model to capture these mappings and make them available for other groups to use.

For example, one of the main driving use cases for Monarch is to address the need to establish links between phenotypes identified in model organisms (e.g., mice, fruit flies, rats, yeast, worms, and zebrafish) and phenotypes identified in humans. Unsurprisingly, the vocabularies used to describe clinical observations and those used to describe model organisms are quite different. Clinical data often refer to "side effects" and "symptoms," whereas model organism data typically refer to "traits" or "phenotypes." In designing Biolink Model, subject matter experts from a variety of disciplines have reconciled these concepts in the "biolink:PhenotypicFeature" class. This makes it possible to query across multiple resources that use multiple terminologies and identifiers and find relevant results.

Illuminating the druggable genome

Illuminating the druggable genome (IDG) aims to identify protein drug targets by developing tools to search, display, and distribute information on these proteins to the biomedical community, whereas supporting research that helps scientists understand how these proteins function. The Illuminating the Druggable Genome Knowledge Graph (KG-IDG) was created to use graph-based machine learning to predict links between drugs and potential targets, in order to identify proteins that are promising drug targets and drugs that are promising repurposing candidates. The generation of this KG relies on Biolink Model to provide a "biolink:Protein" class with mappings to equivalent classes in UniProt, Ensembl, and the Protein Ontology Community. This step ensures that tooling used to identify links among these different protein sources can interrogate them using the same language and same hierarchical data representation. Similarly, KG-IDG uses the "biolink:InformationContentEntity" grouping class to reason over many diverse sources of biomedical attribution, including clinical trials, books, and journal articles. KG-IDG is able to reuse the "biolink:InformationContentEntity" hierarchy in Biolink Model to be specific about the attribution stored in the KG and also reason over the attribution using the higher-level grouping classes, without creating another KG-IDG-specific schema.

Additional applications

Translator, Monarch, and KG-IDG incorporate a broad spectrum of data from a variety of sources, with each source modeling their data using different approaches, independent identifier systems, and heterogeneous data representations. Biolink Model provides the semantic harmonization required to integrate these disparate data sources.

A growing number of other projects also consult and reuse components of the Biolink Model in designing their models. For example, the Alliance of Genome Resources[24] imports some Biolink Model components, even though they do not use the entirety of Biolink Model. Other initiatives that rely on Biolink Model for data and knowledge harmonization include KG-COVID-19[10] and KG-Microbe.[25]

DISCUSSION

The success of Biolink Model can be attributed to its community—biologists, clinicians, data curators, developers, subject matter experts, and ontologists—all of whom have contributed their requirements, perspectives, and expertise to help build a flexible semantic data model. Biolink Model is under continual development, with frequent releases and a publicly accessible issue tracker on GitHub. To ensure sustained development of the model, we invite the biomedical community to contribute via GitHub pull requests and use the issue tracker to suggest new features, report problems, or ask questions (see Supplemental Resources within Supplementary Materials for links to the GitHub repository for Biolink Model, documentation, and other relevant resources).

Biolink Model provides a blueprint to harmonize existing data sources and accelerate the development of new knowledge by leveraging a multitude of domain and technical expertise, captured in a variety of ontologies and existing models (via semantic mappings), within a single modeling framework that is easy to read, write, reuse, and distribute. Moreover, Biolink Model is grounded in semantic web technologies (characterized by classes and slots with their own IRIs, SKOS mappings to existing ontologies, descriptions, identifier prefixes, and domain and range constraints) and captures biomedical expertise as a computable knowledge artifact that can be read and interpreted by both machines and humans. Importantly, KGs that implement Biolink Model immediately gain access to the frameworks and tools developed by a variety of projects that use the model, as well as a platform to connect any Biolink Model–compliant KG to other Translator biomedical KGs.

Because Biolink Model is platform-agnostic, open-source, and publicly accessible, and because it can be translated into a variety of data modeling formats, it encourages people from different backgrounds and with different expertise to work together to evolve the model. Most importantly, Biolink Model supports the harmonization of KGs and underlying data sources in a manner that adheres to FAIR principles[26] and facilitates applications across a broad spectrum of biomedical use cases, thereby democratizing and accelerating translational science.

ACKNOWLEDGMENTS

The authors are grateful to members of the Publications Committees at the National Center for Advancing Translational Sciences, the National Institute of Environmental Health Sciences, and the National Institute on Aging for their review and approval of the manuscript for publication. Moreover, the authors are appreciative of the unwavering leadership and support provided by the Extramural Leadership Team and the Intramural Research Program at National Center for Advancing Translational Sciences (NCATS).

CONFLICTS OF INTEREST

The authors declared no competing interests for this work.

ORCID

Deepak R. Unni https://orcid.org/0000-0002-3583-7340
Sierra A. T. Moxon https://orcid.org/0000-0002-8719-7760
Michael Bada https://orcid.org/0000-0003-3366-4738
Matthew Brush https://orcid.org/0000-0002-1048-5019
Richard Bruskiewich https://orcid.org/0000-0002-4447-5978
Paul A. Clemons https://orcid.org/0000-0002-1800-5112
Vlado Dancik https://orcid.org/0000-0002-5970-6660
Michel Dumontier https://orcid.org/0000-0003-4727-9435
Karamarie Fecho https://orcid.org/0000-0002-6704-9306
Gustavo Glusman https://orcid.org/0000-0001-8060-5955
Jennifer J. Hadlock https://orcid.org/0000-0001-6103-7606
Nomi L. Harris https://orcid.org/0000-0001-6315-3707
Arpita Joshi https://orcid.org/0000-0002-7334-9671
Tim Putman https://orcid.org/0000-0002-4291-0737
Guangrong Qin https://orcid.org/0000-0001-8836-1246
Stephen A. Ramsey https://orcid.org/0000-0002-2168-5403
Kent A. Shefchek https://orcid.org/0000-0001-6439-2224
Harold Solbrig https://orcid.org/0000-0002-5928-3071
Karthik Soman https://orcid.org/0000-0002-3490-9306
Anne E. Thessen https://orcid.org/0000-0002-2908-3327
Melissa A. Haendel https://orcid.org/0000-0001-9114-8737
Chris Bizon https://orcid.org/0000-0002-9491-7674
Christopher J. Mungall https://orcid.org/0000-0002-6601-2165

REFERENCES

1. Bisiani R, Shapiro SC. *Encyclopedia of Artificial Intelligence*. Beam search. Wiley; 1987.

2. National Physical Laboratory. Symposium. International Conference on Machine Translation of Languages and Applied Language Analysis. National Physical Laboratory; 1961. Accessed March 03, 2022. https://market.android.com/details?id=book-dTTawQEACAAJ
3. Vrandečić D, Krötzsch M. *Wikidata: A Free Collaborative Knowledge Base*; 2014. Accessed March 03, 2022. https://ai.google/research/pubs/pub42240
4. Vasilakes JA, Rizvi R, Zhang R. Annotated Semantic Predications from SemMedDB; 2018. Accessed March 03, 2022. https://conservancy.umn.edu/handle/11299/194965
5. Himmelstein DS, Lizee A, Hessler C, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*. 2017;22:6. doi:10.7554/eLife.26726
6. Hettne KM, Thompson M, van Haagen HHHBM, et al. The implicitome: a resource for rationalizing gene-disease associations. *PLoS One*. 2016;11(2):e0149621. doi:10.1371/journal.pone.0149621
7. Shefchek KA, Harris NL, Gargano M, et al. The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res*. 2020;48(D1):D704-D715. doi:10.1093/nar/gkz997
8. Putman TE, Lelong S, Burgstaller-Muehlbacher S, et al. WikiGenomes: an open web application for community consumption and curation of gene annotation data in Wikidata. *Database*. Narnia. 2017. doi:10.1093/database/bax025/3084697
9. Nelson CA, Acuna AU, Paul AM, et al. Knowledge network embedding of transcriptomic data from spaceflown mice uncovers signs and symptoms associated with terrestrial diseases. *Life*. 2021;11(1):42. doi:10.3390/life11010042
10. Reese JT, Unni D, Callahan TJ, et al. KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19 response. *Patterns (N Y)*. 2021;2(1):100155. doi:10.1016/j.patter.2020.100155
11. Hogan A, Blomqvist E, Cochez M, et al. Knowledge graphs. *arXiv. [cs.AI]*; 2020. Accessed March 03, 2022. http://arxiv.org/abs/2003.02320
12. Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res*. 2001;11(8):1425-1433. doi:10.1101/gr.180801
13. Mungall CJ, Dietze H, Osumi-Sutherland D. Use of OWL within the Gene Ontology. In: Keet M, Tamma V, eds. *Proceedings of the 11th International Workshop on OWL: Experiences and Directions (OWLED 2014)*. Riva del Garda, Italy, October 17-18, 2014; 2014:25-36. Accessed March 03, 2022. http://ceur-ws.org/Vol-1265/owled2014_submission_5.pdf
14. Alshahrani M, Samothrakis S, Fasli M. Word mover's distance for affect detection. *2017 International Conference on the Frontiers and Advances in Data Science (FADS)*. IEEE; 2017:18-23. doi:10.1109/FADS.2017.8253186
15. Biomedical Data Translator Consortium. Toward a universal biomedical data translator. *Clin Transl Sci*. 2019;12(2):86-90. doi:10.1111/cts.12591
16. Jackson R, Matentzoglu N, Overton JA, et al. OBO foundry in 2021: operationalizing open data principles to evaluate ontologies. *Database*. 2021;26:2021. doi:10.1093/database/baab069
17. Dumontier M, Baker CJ, Baran J, et al. The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *J Biomed Semantics*. 2014;5(1):14. doi:10.1186/2041-1480-5-14
18. Samuel J. Collaborative approach to developing a multilingual ontology: a case study of Wikidata. *Metadata and Semantic Research*. 2017;167-172. doi:10.1007/978-3-319-70863-8_16
19. McMurry JA, Juty N, Blomberg N, et al. Identifiers for the 21st century: how to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLoS Biol*. 2017;15(6):e2001414. doi:10.1371/journal.pbio.2001414
20. Vasilevsky N, Essaid S, Matentzoglu N, et al. Mondo disease ontology: harmonizing disease concepts across the world. CEUR Workshop Proceedings CEUR-WS; 2020. Accessed March 03, 2022. http://ceur-ws.org/Vol-2807/abstractY.pdf
21. OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition). Accessed March 03, 2022. https://www.w3.org/TR/owl2-syntax/
22. Fecho K, Thessen AT, Baranzini SE, et al. and The Biomedical Data Translator Consortium. Progress toward a universal biomedical data translator. *Clin Transl Sci*. doi: 10.1111/cts.13301
23. Austin CP, Colvis CM, Southall NT. Deconstructing the translational tower of babel. *Clin Transl Sci*. 2019;12(2):85. doi:10.1111/cts.12595
24. Alliance of Genome Resources Consortium. Harmonizing model organism data in the Alliance of Genome Resources. *Genetics*. 2022;220(4). doi:10.1093/genetics/iyac022
25. Joachimiak MP, Hegde H, Duncan WD, et al. KG-microbe: a reference Knowledge-Graph and platform for harmonized microbial information; 2021. Accessed March 03, 2022. http://ceur-ws.org/Vol-3073/paper19.pdf
26. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data*. 2016;15(3):160018. doi:10.1038/sdata.2016.18

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Unni DR, Moxon SAT, Bada M, et al. Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. *Clin Transl Sci*. 2022;00:1-8. doi:10.1111/cts.13302