Wikidata as a FAIR knowledge graph for the life sciences

Andra Waagmeester1,*, Gregory Stupp2,*, Sebastian Burgstaller-Muehlbacher2,†, Benjamin M. Good2, Malachi Griffith3, Obi Griffith3, Kristina Hanspers4, Henning Hermjakob5, Kevin Hybiske6, Sarah M. Keating5, Magnus Manske7, Michael Mayers2, Elvira Mitraka8, Alexander R. Pico4, Timothy Putman2, Anders Riutta4, Núria Queralt-Rosinach2, Lynn M. Schriml8, Denise Slenter9, Ginger Tsueng2, Roger Tu2, Egon Willighagen9, Chunlei Wu2, Andrew I. Su2,§

1 Micelio, Antwerp 2180, Belgium
2 The Scripps Research Institute, La Jolla, CA, USA
3 McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA
4 Institute of Data Science and Biotechnology, Gladstone Institutes, San Francisco, CA, USA
5 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, CB10 1SD, Hinxton, United Kingdom
6 Division of Allergy and Infectious Diseases, Department of Medicine, University of Washington, Seattle, WA, USA
7 Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK
8 University of Maryland School of Medicine, Baltimore, Maryland, USA.
9 Department of Bioinformatics-BiGCaT, NUTRIM, Maastricht University, 6229 ER Maastricht, The Netherlands

(*) equal contributions
(§) Correspondence: asu@scripps.edu
(¶) Present address: Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna, Medical University Vienna, Austria

ORCID identifiers:
Andra Waagmeester, 0000-0001-9773-4008
Gregory Stupp, 0000-0002-0644-7212
Sebastian Burgstaller-Muehlbacher, 0000-0003-4640-3510
Benjamin M. Good, 0000-0002-7334-7852
Malachi Griffith, 0000-0002-6388-446X
Obi Griffith, 0000-0002-0843-4271
Kristina Hanspers 0000-0001-5410-599X
Henning Hermjakob, 0000-0001-8479-0262
Kevin Hybiske, 0000-0002-2967-3079
Sarah M. Keating, 0000-0002-3356-3542
Magnus Manske, 0000-0001-5916-0947
Michael Mayers, 0000-0002-7792-0150

Elvira Mitraka, 0000-0003-0719-3485
Alexander R. Pico 0000-0001-5706-2163
Timothy Putman, 0000-0002-4291-0737
Anders Riutta 0000-0002-4693-0591
Núria Queralt-Rosinach, 0000-0003-0169-8159
Lynn M. Schriml, 0000-0001-8910-9851
Denise Slenter, 0000-0001-8449-1318
Ginger Tsueng, 0000-0001-9536-9115
Roger Tu, 0000-0002-7899-1604
Egon Willighagen, 0000-0001-7542-0286
Chunlei Wu, 0000-0002-2629-6124
Andrew I. Su, 0000-0002-9859-4104
Abstract

Wikidata is a community-maintained knowledge base that epitomizes the FAIR principles of Findability, Accessibility, Interoperability, and Reusability. Here, we describe the breadth and depth of biomedical knowledge contained within Wikidata, assembled from primary knowledge repositories on genomics, proteomics, genetic variants, pathways, chemical compounds, and diseases. We built a collection of open-source tools that simplify the addition and synchronization of Wikidata with source databases. We furthermore demonstrate several use cases of how the continuously updated, crowd-contributed knowledge in Wikidata can be mined. These use cases cover a diverse cross section of biomedical analyses, from crowdsourced curation of biomedical ontologies, to phenotype-based diagnosis of disease, to drug repurposing.

Introduction

Integrating data and knowledge is a formidable challenge in biomedical research. Although new scientific findings are being discovered at a rapid pace, a large proportion of that knowledge is either locked in data silos (where integration is hindered by differing nomenclature, data models, and licensing terms) [1], or even worse, locked away in free-text. The lack of an integrated and structured version of biomedical knowledge hinders efficient querying or mining of that information, a limitation that prevents the full utilization of our accumulated scientific knowledge.

Recently, there has been a growing emphasis within the scientific community to ensure all scientific data are FAIR – Findable, Accessible, Interoperable, and Reusable – and there is a growing consensus around a concrete set of principles to ensure FAIRness [1,2]. Widespread implementation of these principles would greatly advance open data efforts to build a rich and heterogeneous network of scientific knowledge. That knowledge network could, in turn, be the foundation for many computational tools, applications and analyses.

Most data and knowledge integration initiatives fall on either end of a spectrum. At one end, centralized efforts seek to bring all knowledge sources into a single database instance (e.g., [3]). This approach has the advantage of data alignment according to a common data model and of enabling high performance queries. However, centralized resources are very difficult and expensive to maintain and expand [4,5], in large part because of limited bandwidth and resources of the technical team and the bottlenecks that introduces.

At the other end of the spectrum, distributed approaches to data integration leave in place a broad landscape of individual resources, focusing on technical infrastructure to query and integrate across them for each query. These approaches lower the barriers to adding new data by enabling anyone to publish data by following community standards. However, performance is often an issue when each query must be sent to many individual databases, and the performance of the system as a whole is highly dependent on the stability and performance of each individual component. In addition, data
integration requires harmonizing the differences in the data models and data formats between resources, a process that can often require significant skill and effort.

Here we explore the use of Wikidata (https://www.wikidata.org) [6] as a platform for knowledge integration in the life sciences. Wikidata is an openly-accessible knowledge base that is editable by anyone. Like its sister project Wikipedia, the scope of Wikidata is nearly boundless, with items on topics as diverse as books, actors, historical events, and galaxies. Unlike Wikipedia, Wikidata focuses on representing knowledge in a structured format instead of primarily free text. As of September 2019, Wikidata’s knowledge graph included over 750 million statements on 61 million items [7]. Wikidata also became the first Wikimedia project that surpassed one billion edits, achieved by its community of 20 thousand active users and 80 active computational ‘bots’. Since its inception in 2012, Wikidata has a proven track record for leveraging the crowdsourced efforts of engaged users in building a massive knowledge graph [8]. Wikidata is run by the Wikimedia Foundation (https://wikimediafoundation.org), an organization that has a long track record of developing and maintaining web applications at scale.

As a knowledge integration platform, Wikidata combines several of the key strengths of the centralized and distributed approaches. A large portion of the Wikidata knowledge graph is based on the automated imports of large structured databases via Wikidata bots, thereby breaking down the walls of existing data silos. Since Wikidata is also based on a community-editing model, it harnesses the distributed efforts of a worldwide community of contributors. Anyone is empowered to add new statements, ranging from individual facts to large-scale data imports. Finally, all knowledge in Wikidata is queryable through a SPARQL query interface [9], which enables distributed queries across other Linked Data resources.

In previous work, we seeded Wikidata with content from public and authoritative resources on structured knowledge on genes and proteins [10] and chemical compounds [11]. Here, we describe progress on expanding and enriching the biomedical knowledge graph within Wikidata, both by our team and by others in the community [12]. We also describe several representative use cases on how Wikidata can enable new analyses and improve the efficiency of research. Finally, we discuss how researchers can contribute to this effort to build a continuously-updated and community-maintained knowledge graph that epitomizes the FAIR principles.

Results

The Wikidata Biomedical Knowledge Graph

The original effort behind this work focused on creating and annotating Wikidata items for human and mouse genes and proteins [10], and was subsequently expanded to include microbial reference genomes from NCBI RefSeq [13]. Since then, the Wikidata community (including our team) has significantly expanded the depth and breadth of biological information within Wikidata, resulting in a rich, heterogeneous knowledge graph (Figure 1). Some of the key new data types and resources are described below.
**Figure 1. A class-level diagram of the Wikidata knowledge graph for biomedical entities.** Each box represents one type of biomedical entity. The header displays the name of that entity type, as well as the count of Wikidata items of that type. The lower portion of each box displays a partial listing of attributes about each entity type, together with the count of the number of items with that attribute. Edges between boxes represent the number of Wikidata statements corresponding to each combination of subject type, predicate, and object type. For clarity, edges for reciprocal relationships (e.g., “has part” and “part of”) are combined into a single edge. All counts of Wikidata items are current as of September 2019. Data are generated using the code in https://github.com/SuLab/genewikiworld.

**Genes and proteins.** Wikidata contains items for over 1.1 million genes and 940 thousand proteins from 201 unique taxa. Annotation data on genes and proteins come from several key databases including NCBI Gene [14], Ensembl [15], UniProt [16], InterPro [17], and the Protein Data Bank (PDB) [18]. These annotations include information on protein families, gene functions, protein domains, genomic location, and orthologs, as well as links to related compounds, diseases, and variants.

**Genetic variants.** Annotations on genetic variants are primarily drawn from CIViC (http://www.civicdb.org), an open and community-curated database of cancer variants [19]. Variants are annotated with their relevance to disease predisposition, diagnosis, prognosis, and drug efficacy. Wikidata currently contains 1502 items corresponding to human genetic variants, focused on those with a clear clinical or therapeutic relevance.

**Chemical compounds including drugs.** Wikidata has items for over 150 thousand chemical compounds, including over 3500 items which are specifically designated as medications. Compound attributes are drawn from a diverse set of databases, including PubChem [20], RxNorm [21], IUPHAR Guide to Pharmacology [22–24], NDF-RT [25], and LIPID MAPS [26]. These items typically contain statements describing chemical structure and key physicochemical properties, and links to databases with experimental data (MassBank [27,28], PDB Ligand [29], etc.) and toxicological information (EPA CompTox Dashboard [30]). Additionally, these items contain links to compound classes, disease indications, pharmaceutical products, and protein targets.

**Pathways.** Wikidata has items for almost three thousand human biological pathways, primarily from two established public pathway repositories: Reactome [31] and WikiPathways [32]. The full details of the different pathways remain with the respective primary sources. Our bots enter data for Wikidata properties such as pathway name, identifier, organism, and the list of component genes, proteins, and chemical compounds. Properties for contributing authors (via ORCID properties [33]), descriptions and ontology annotations are also being added for Wikidata pathway entries.

**Diseases.** Wikidata has items for over 16 thousand diseases, the majority of which were created based on imports from the Human Disease Ontology [34], with additional disease terms added from the Monarch Disease Ontology [3]. Disease attributes include medical classifications, symptoms, relevant drugs, as well as subclass relationships to higher-level disease categories. In instances where the Human Disease Ontology specifies a related anatomic region and/or a causative organism (for infectious diseases), corresponding statements are also added.
References. Whenever practical, the provenance of each statement added to Wikidata was also added in a structured format. References are part of the core data model for a Wikidata statement. References can either cite the primary resource from which the statement was retrieved (including details like version number of the resource), or they can link to a Wikidata item corresponding to a publication as provided by a primary resource (as an extension of the WikiCite project [35]), or both.

Bot automation

To programmatically upload biomedical knowledge to Wikidata, we developed a series of computer programs, or bots. Bot development began by reaching a consensus on data modeling with the Wikidata community, particularly the Molecular Biology WikiProject [36]. We then coded each bot to perform data retrieval from a primary resource, data transformation and normalization, and then data upload via the Wikidata application programming interface (API).

We generalized the common code modules into a Python library, called Wikidata Integrator (WDI), to simplify the process of creating Wikidata bots [37]. Relative to accessing the API directly, WDI has convenient features that improve the bot development experience. These features include the creation of items for scientific articles as references, basic detection of data model conflicts, automated detection of items needing update, detailed logging and error handling, and detection and preservation of conflicting human edits.

Just as important as the initial data upload is the synchronization of updates between the primary sources and Wikidata. We utilized Jenkins, an open-source automation server, to automate all our Wikidata bots. This system allows for flexible scheduling, job tracking, dependency management, and automated logging and notification. Bots are either run on a predefined schedule (for continuously updated resources) or when new versions of original databases are released.

Applications

Identifier Translation

Translating between identifiers from different databases is one of the most common operations in bioinformatics analyses. Unfortunately, these translations are most often done by bespoke scripts and based on entity-specific mapping tables. These translation scripts are repetitively and redundantly written across our community and are rarely kept up to date.

An identifier translation service is a simple and straightforward application of the biomedical content in Wikidata. Based on mapping tables that have been imported, Wikidata items can be mapped to databases that are both widely- and rarely-used in the life sciences community. Because all these mappings are stored in a centralized database and use a systematic data model, generic and reusable translation scripts can easily be written (Figure 2). These scripts can be used as a foundation for more complex Wikidata queries, or the results can be downloaded and used as part of larger scripts or analyses.
There are a number of other tools that are also aimed at solving the identifier translation use case, including the BioThings APIs [38], BridgeDb [39], BioMart [40], UMLS [41], and NCI Thesaurus [42]. Relative to these tools, Wikidata distinguishes itself with a unique combination of the following:

- an almost limitless scope including all entities in biology, chemistry, and medicine;
- a data model that can represent exact, broader, and narrow matches between items in different identifier namespaces (beyond semantically imprecise "cross-references");
- programmatic access through web services with a track record of high-performance and high-availability

Moreover, Wikidata is also unique as it is the only tool that allows real-time community editing. So while Wikidata is certainly not complete with respect to identifier mappings, it can be continually improved independent of any centralized effort or curation authority.

**Figure 2. Generalizable SPARQL template for identifier translation.** This simple example shows how identifiers of any biological type can easily be translated using SPARQL queries. These queries operate on Wikidata properties for gene symbols (wdt:P353) and Entrez Gene IDs (wdt:P351) (top), and RxNorm concept IDs (wdt:P3345) and NDF-RT IDs (wdt:P2115) (bottom). These queries can be submitted to the Wikidata Query Service (WDQS; https://query.wikidata.org/) to get real-time results from Wikidata data. RELatively simple extensions of these queries can also be added to filter mappings based on the statement references and/or qualifiers. A full list of Wikidata properties can be found at [43]. Note that for translating a large number of identifiers, it is often more efficient to perform a SPARQL query to retrieve all mappings and then perform additional filtering locally.

**Integrative Queries**

Wikidata contains a much broader set of information than just identifier cross-references. Having biomedical data in one centralized data resource facilitates powerful integrative queries that span multiple domain areas and data sources. Performing these integrative queries through Wikidata obviates the need to perform many time-consuming and error-prone data integration steps.

As an example, consider a pulmonologist who is interested in identifying candidate chemical compounds for testing in disease models (schematically illustrated in Figure 3). She may start by identifying genes with a genetic association to any respiratory disease, with a particular interest in genes that encode membrane-bound proteins (for ease in cell sorting). She may then look for chemical
compounds that either directly inhibit those proteins, or finding none, compounds that inhibit another protein in the same pathway. Because she has collaborators with relevant expertise, she may specifically filter for proteins containing a serine-threonine kinase domain.

Almost any competent informatician can perform the query described above by integrating cell localization data from Gene Ontology annotations, genetic associations from GWAS Catalog, disease subclass relationships from the Human Disease Ontology, pathway data from WikiPathways and Reactome, compound targets from the IUPHAR Guide to Pharmacology, and protein domain information from InterPro. However, actually performing this data integration is a time-consuming and error-prone process. At the time of publication of this manuscript, this Wikidata query completed in less than 10 seconds and reported 31 unique compounds. Importantly, the results of that query will always be up-to-date with the latest information in Wikidata.

This query, and other example SPARQL queries that take advantage of the rich, heterogeneous knowledge network in Wikidata are available at [https://www.wikidata.org/wiki/User:ProteinBoxBot/SPARQL_Examples](https://www.wikidata.org/wiki/User:ProteinBoxBot/SPARQL_Examples). That page additionally demonstrates federated SPARQL queries that perform complex queries across other biomedical SPARQL endpoints. Federated queries are useful for accessing data that cannot be included in Wikidata directly due to limitations in size, scope, or licensing.

![Figure 3. A representative SPARQL query that integrates data from multiple data resources and annotation types.](https://example.com/sparql_query)

This query incorporates data on genetic associations to disease, Gene Ontology annotations for cellular compartment, protein target information for compounds, pathway data, and protein domain information. More context is provided in the text. Real-time query results can be viewed at [https://w.wiki/6pZ](https://w.wiki/6pZ).

**Crowdsourced Curation**

Ontologies are essential resources for structuring biomedical knowledge. However, even after the initial effort in creating an ontology is finalized, significant resources must be devoted to maintenance and further development. These tasks include cataloging cross references to other ontologies and vocabularies, and modifying the ontology as current knowledge evolves. Community curation has been explored in a variety of tasks in ontology curation and annotation (e.g., [13,44–47]). While community
curation offers the potential of distributing these responsibilities over a wider set of scientists, it also has the potential to introduce errors and inconsistencies.

Here, we examined how a crowd-based curation model through Wikidata works in practice. We designed a system to monitor, filter, and prioritize changes made by Wikidata contributors to items in the Human Disease Ontology. We initially seeded Wikidata with disease items from the Disease Ontology (DO) starting in late 2015. Beginning in 2018, we compared the disease data in Wikidata to the most current DO release on a monthly basis.

In our first comparison between Wikidata and the official DO release, we found that Wikidata users added a total of 2030 new cross references to GARD [48] and MeSH [49]. Each cross reference was manually reviewed by DO curators, and 98.9% of these mappings were deemed correct and therefore added to the ensuing DO release. Each subsequent monthly report included a smaller number of added cross references to GARD and MeSH, as well as ORDO [50], and OMIM [51,52], and these entries were incorporated after expert review at a high approval rate (>90%). Wikidata users also suggested numerous refinements to the ontology structure, including changes to the subclass relationships and the addition of new disease terms. While these structural changes were rarely incorporated into DO releases with no modifications, they often prompted further review and refinement by DO curators in specific subsections of the ontology.

The Wikidata crowdsourcing curation model is generalizable to any other external resource that is automatically synced to Wikidata. The code to detect changes and assemble reports is tracked online [53] and can easily be adapted to other domain areas. This approach offers a novel solution for integrating new knowledge into a biomedical ontology through distributed crowdsourcing while preserving control over the expert curation process. Incorporation into Wikidata also enhances exposure and visibility of the resource by engaging a broader community of users and curators.

Interactive Pathway Pages

In addition to its use as a repository for data, we explored the use of Wikidata as a primary access and visualization endpoint for pathway data. We used Scholia, a web app for displaying scholarly profiles for a variety of Wikidata entries, including individual researchers, research topics, chemicals, and proteins [11]. Scholia provides a more user-friendly view of Wikidata content with context and interactivity that is tailored to the entity type.

We contributed a Scholia profile template specifically for biological pathways [54,55]. In addition to essential items such as title and description, these pathway pages include an interactive view of the pathway diagram collectively drawn by contributing authors. The WikiPathways identifier property in Wikidata informs the Scholia template to source a pathway-viewer widget from Toolforge [56] that in turn retrieves the corresponding interactive pathway image. Embedded into the Scholia pathway page, the widget provides pan and zoom, plus links to gene, protein and chemical Scholia pages for every clickable molecule on the pathway diagram (see for example [57]). Each pathway page also includes information about the pathway authors. The Scholia template also generates a participants table that shows the genes, proteins, metabolites, and chemical compounds that play a role in the pathway, as well as citation information in both tabular and chart formats.
With Scholia template views of Wikidata, we are able to generate interactive pathway pages with comparable content and functionality to that of dedicated pathway databases. Wikidata provides a powerful interface to access these biological pathway data in the context of other biomedical knowledge, and Scholia templates provide rich, dynamic views of Wikidata that are relatively simple to develop and maintain.

Phenotype-based disease diagnosis

Phenomizer is a web application that suggests clinical diagnoses based on an array of patient phenotypes. Phenomizer takes as input a list of phenotypes (using the Human Phenotype Ontology (HPO) [58]) and an association file between phenotypes and diseases, and the Phenomizer algorithm suggests disease diagnoses based on semantic similarity [59]. Here, we studied whether phenotype-disease associations from Wikidata could improve Phenomizer’s ability to make differential diagnoses for certain sets of phenotypes. We modified the Phenomizer codebase to accept arbitrary inputs and to be able to run from the command line [60] and also wrote a script to extract and incorporate the phenotype-disease annotations in Wikidata [61].

As of September 2019, there were 273 phenotype-disease associations in Wikidata that were not in the HPO’s annotation file (which contained a total of 172,760 associations). Based on parallel biocuration work by our team, many of these new associations were related to the disease Congenital Disorder of Deglycosylation (CDDG; also known as NGLY-1 deficiency). To see if the Wikidata-sourced annotations improved the ability of Phenomizer to diagnose CDDG, we ran our modified version using the phenotypes taken from a publication describing two siblings with suspected cases of CDDG [62]. Using these phenotypes and the annotation file supplemented with Wikidata-derived associations, Phenomizer returned a much stronger semantic similarity to CDDG relative to the HPO annotation file alone (Figure 4). Analyses with the combined annotation file reported CDDG as the top result for each of the past 14 releases of the HPO annotation file, whereas CDDG was never the top result when run without the Wikidata-derived annotations. This result demonstrated an example scenario in which Wikidata-derived annotations could be a useful complement to expert curation.

Figure 4. Phenomizer analysis of suspected cases of CDDG. Clinical phenotypes from two cases of suspected CDDG patients were extracted from a published case report [62]. These phenotypes were run through the Phenomizer tool using phenotype-disease annotations from HPO alone, or from a combination of HPO and Wikidata. The semantic similarity score for CDDG is reported on the y-axis.
Drug Repurposing

The mining of graphs for latent edges has been an area of interest in a variety of contexts from predicting friend relationships in social media platforms to suggesting movies based on past viewing history. A number of groups have explored the mining of knowledge graphs to reveal biomedical insights, with the open source Rephetio effort for drug repurposing as one example [63]. Rephetio uses logistic regression, with features based on graph metapaths, to predict drug repurposing candidates.

The knowledge graph that served as the foundation for Rephetio was manually assembled from many different resources into a heterogeneous knowledge network. Here, we explored whether the Rephetio algorithm could successfully predict drug indications on the Wikidata knowledge graph. Based on the class diagram in Figure 1, we extracted a biomedically-focused subgraph of Wikidata with 19 node types and 41 edge types. We performed five-fold cross validation on drug indications within Wikidata and found that Rephetio substantially enriched for the true indications in the hold-out set. We then downloaded historical Wikidata versions from 2017 and 2018, and observed marked improvements in performance over time (Figure 6). We also performed this analysis using an external test set based on Drug Central, which showed a similar improvement in Rephetio results over time (Supplemental Figure 1).

This analysis demonstrates the value of a community-maintained, centralized knowledge base to which many researchers are contributing. It suggests that scientific analyses based on Wikidata may continually improve irrespective of any changes to the underlying algorithms, but simply based on progress in curating knowledge through the distributed, and largely uncoordinated efforts of the Wikidata community.

Figure 5. Drug repurposing using the Wikidata knowledge graph. We analyzed three snapshots of Wikidata using Rephetio, a graph-based algorithm for predicting drug repurposing candidates [63]. We evaluated the performance of the Rephetio algorithm on three historical versions of the Wikidata knowledge graph, quantified based on the area under the receiver operator characteristic curve (AUC). This analysis demonstrated that the performance of Rephetio in drug repurposing improved over time based only on improvements to the underlying knowledge graph. Details of this analysis can be found at https://github.com/SuLab/WD-rephetio-analysis.
Supplemental Figure 1. Drug repurposing using the Wikidata knowledge graph, evaluated using an external test set. Whereas the analysis in Figure 5 was based on a cross-validation of indications that were present in Wikidata, we also ran our time-resolved analysis using an external gold standard set of indications from Drug Central [64].

Discussion

We believe that Wikidata is among the most FAIR biomedical resources available (a view that is also shared among some funding bodies [65]).

- **Findable**: Wikidata items are assigned globally unique identifiers with direct cross-links into the massive online ecosystem of Wikipedias. Wikidata also has broad visibility within the Linked Data community and listed in the life science registries FAIRsharing [66] and Identifiers.org [67]. Wikidata has already attracted a robust, global community of contributors and consumers.

- **Accessible**: Wikidata provides access to its underlying knowledge graph via both an online graphical user interface and an API, and access includes both read- and write-privileges. Wikidata also provides database dumps at least weekly [68], ensuring the long-term accessibility of the Wikidata knowledge graph independent of the organization and web application.

- **Interoperable**: Wikidata items are extensively cross-linked to other biomedical resources using Universal Resource Identifiers (URIs), which unambiguously anchor these concepts in the Linked Open Data cloud [69]. Wikidata is also available in many standard formats in computer programming and knowledge management, including JSON, XML, and RDF.

- **Reusable**: Data provenance is directly tracked in the reference section of the Wikidata statement model. The Wikidata knowledge graph is released under the Creative Commons Zero (CC0) Public Domain Declaration, which explicitly declares that there are no restrictions on downstream reuse and redistribution [70].

The open data licensing of Wikidata is particularly notable. The use of data licenses in biomedical research has rapidly proliferated, presumably in an effort to protect intellectual property and/or justify long-term grant funding (e.g. [71]). However, even seemingly innocuous license terms (like requirements for attribution) still impose legal requirements and therefore expose consumers to legal liability. This liability is especially problematic for data integration efforts, in which the license terms of
all resources (dozens or hundreds or more) must be independently tracked and satisfied (a phenomenon referred to as "license stacking"). Because it is released under CC0, Wikidata can be freely and openly used in any other resource without any restriction. This freedom greatly simplifies and encourages downstream use.

In addition to simplifying data licensing, Wikidata offers significant advantages in centralizing the data harmonization process. Consider the use case of trying to get a comprehensive list of disease indications for the drug bupropion. The National Drug File - Reference Terminology (NDF-RT) reported that bupropion may treat nicotine dependence and attention deficit hyperactivity disorder, the Inxight database listed major depressive disorder, and the FDA Adverse Event Reporting System (FAERS) listed anxiety and bipolar disorder. While no single database listed all these indications, Wikidata provided an integrated view that enabled seamless query and access across resources. Integrating drug indication data from these individual data resources was not a trivial process. Both Inxight and NDF-RT mint their own identifiers for both drugs and diseases. FAERS uses Medical Dictionary for Regulatory Activities (MedDRA) names for diseases and free-text names for drugs [72]. By harmonizing and integrating all resources in the context of Wikidata, we ensure that those data are immediately usable by others without having to repeat the normalization process. Moreover, by harmonizing data at the time of data loading, consumers of that data do not need to perform the repetitive and redundant work at the point of querying and analysis.

As the biomedical data within Wikidata continues to grow, we believe that its unencumbered use will spur the development of many new innovative tools and analyses. These innovations will undoubtedly include the machine learning-based mining of the knowledge graph to predict new relationships (also referred to as knowledge graph reasoning [73–75]).

For those who subscribe to this vision for cultivating a FAIR and open graph of biomedical knowledge, there are two simple ways to contribute to Wikidata. First, owners of data resources can release their data using the CC0 declaration. Because Wikidata is released under CC0, it also means that all data imported in Wikidata must also use CC0-compatible terms (e.g., be in the public domain). For resources that currently use a restrictive data license primarily for the purposes of enforcing attribution or citation, we encourage the transition to "CC0 (+BY)", a model that "[pairs] a permissive license with a strong moral entreaty" [76]. For resources that must retain data license restrictions, consider releasing a subset of data or older versions of data using CC0. Many biomedical resources were created under or transitioned to CC0 (in part or in full) in recent years [77], including the Disease Ontology [34], Pfam [78], Bgee [79], WikiPathways [32], Reactome [31], ECO [80], and CIViC [19].

Second, informaticians can contribute to Wikidata by adding the results of data parsing and integration efforts to Wikidata. Currently, the useful lifespan of data integration code typically does not extend beyond the immediate project-specific use. As a result, that same data integration process is likely being done repetitively and redundantly by other informaticians elsewhere. If every informatician contributed the output of their effort to Wikidata, the resulting knowledge graph would be far more useful than the stand-alone contribution of any single individual, and it would continually improve in both breadth and depth over time.
FAIR and open access to the sum total of biomedical knowledge will improve the efficiency of biomedical research. Capturing that information in a centralized knowledge graph is useful for experimental researchers, informatics tool developers and biomedical data scientists. As a continuously-updated and collaboratively-maintained community resource, we believe that Wikidata has made significant strides toward achieving this ambitious goal.

Acknowledgments

The authors thank the thousands of Wikidata contributors for curating knowledge, both directly related and unrelated to this work. The authors also thank the Wikimedia Foundation developers and administrators for maintaining Wikidata as a community resource. This work has been supported by grants from the National Institute for General Medical Science (NIGMS) under awards R01 GM089820 to AS, U54 GM114833 to AS and HH, and R01 GM100039 to AP. MG is supported by the National Human Genome Research Institute (NHGRI) of the NIH under Award Number R00HG007940, the National Cancer Institute under Award Number U24CA237719 and the V Foundation for Cancer Research under Award Number V2018-007. KH was supported by the National Institute of Allergy and Infectious Diseases (NIAID) under award R01 AI126785 Additional support was provided by the National Center for Advancing Translational Sciences (NCATS) under award UL1 TR002550.

Competing interests

The authors have no competing interests.
References


33. Sprague ER. ORCID. J Med Libr Assoc JMLA. 2017 Apr;105(2):207–208. PMCID: PMC5370620


57. Scholia, ACE Inhibitor Pathway [Internet]. Available from: https://tools.wmflabs.org/scholia/pathway/Q29892242


Bayesian ontology querying from Bauer et al. Contribute to SuLab/boqa development by creating an account on GitHub [Internet]. Su Lab; 2018 [cited 2019 Jul 23]. Available from: https://github.com/SuLab/boqa

Incorporate wikidata statements into phenomizer. Contribute to SuLab/Wikidata-phenomizer development by creating an account on GitHub [Internet]. Su Lab; 2019 [cited 2019 Jul 23]. Available from: https://github.com/SuLab/Wikidata-phenomizer


Creative Commons — CC0 1.0 Universal [Internet]. [cited 2019 Aug 8]. Available from: https://creativecommons.org/publicdomain/zero/1.0/


76. CC0 (+BY) – Dan Cohen [Internet]. [cited 2019 Aug 8]. Available from: https://dancohen.org/2013/11/26/cc0-by/

77. FAIRsharing [Internet]. [cited 2019 Jan 25]. Available from: https://fairsharing.org/

