eLIFE elifesciences.org

FEATURE ARTICLE

SCIENCE FORUM

Wikidata as a knowledge graph for the life sciences

Abstract Wikidata is a community-maintained knowledge base that has been assembled from repositories in the fields of genomics, proteomics, genetic variants, pathways, chemical compounds, and diseases, and that adheres to the FAIR principles of findability, accessibility, interoperability and reusability. Here we describe the breadth and depth of the biomedical knowledge contained within Wikidata, and discuss the open-source tools we have built to add information to Wikidata and to synchronize it with source databases. We also demonstrate several use cases for Wikidata, including the crowdsourced curation of biomedical ontologies, phenotype-based diagnosis of disease, and drug repurposing.

ANDRA WAAGMEESTER†, GREGORY STUPP†, SEBASTIAN BURGSTALLER-MUEHLBACHER, BENJAMIN M GOOD, MALACHI GRIFFITH, OBI L GRIFFITH, KRISTINA HANSPERS, HENNING HERMJAKOB, TOBY S HUDSON, KEVIN HYBISKE, SARAH M KEATING, MAGNUS MANSKE, MICHAEL MAYERS, DANIEL MIETCHEN, ELVIRA MITRAKA, ALEXANDER R PICO, TIMOTHY PUTMAN, ANDERS RIUTTA, NURIA QUERALT-ROSINACH, LYNN M SCHRIML, THOMAS SHAFEE, DENISE SLENTER, RALF STEPHAN, KATHERINE THORNTON, GINGER TSUENG, ROGER TU, SABAH UL-HASAN, EGON WILLIGHAGEN, CHUNLEI WU AND ANDREW I SU*

***For correspondence:** asu@scripps.edu

†These authors contributed equally to this work

Competing interests: The authors declare that no competing interests exist.

Funding: See page 12

Reviewing editor: Peter Rodgers, eLife, United Kingdom

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; ***Wilkinson et al., 2016***) or locked away in free-text. The lack of an integrated and structured version of biomedical knowledge hinders efficient querying or mining of that information, thus preventing 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 (***Wilkinson et al., 2019***; ***Wilkinson et al., 2016***). Widespread implementation of these principles would greatly advance efforts by the open-data community 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 multiple knowledge sources into a single database (see, for example, ***Mungall et al., 2017***): this approach has the advantage of data alignment according to a common data model and of enabling high performance queries. However, centralized resources are difficult and expensive to maintain and expand (***Chandras et al., 2009***; ***Gabella et al., 2018***), at least in part because of bottlenecks that are inherent in a centralized design.

At the other end of the spectrum, distributed approaches to data integration result in 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. Moreover, harmonizing differences in data licensing can sometimes be impossible.

Here we explore the use of Wikidata (www.wikidata.org; ***Vrandečić, 2012***; ***Mora-Cantallops et al., 2019***) 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 (tools.wmflabs.org/wikidata-todo/stats.php). Wikidata was also the first project run by the Wikimedia Foundation (which also runs Wikipedia) to have surpassed one billion edits, achieved by a community of 12,000 active users, including 100 active computational 'bots' (***Figure 1—figure supplement 1***).

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, including both domain experts and bot developers. 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 (query.wikidata.org/), which also enables distributed queries across other Linked Data resources.

In previous work, we seeded Wikidata with content from public and authoritative sources of structured knowledge on genes and proteins (***Burgstaller-Muehlbacher et al., 2016***) and chemical compounds (***Willighagen et al., 2018***). Here, we describe progress on expanding and enriching the biomedical knowledge graph within Wikidata, both by our team and by others in the community (***Turki et al., 2019***). We also describe several representative biomedical 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.

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 (***Burgstaller-Muehlbacher et al., 2016***), and was subsequently expanded to include microbial reference genomes from NCBI RefSeq (***Putman et al., 2017***). 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.

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 (***Agarwala et al., 2018***), Ensembl (***Zerbino et al., 2018***), UniProt (***UniProt Consortium, 2019***), InterPro (***Mitchell et al., 2019***), and the Protein Data Bank (***Burley et al., 2019***). 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 (***Griffith et al., 2017***). Variants are annotated with their relevance to disease predisposition, diagnosis, prognosis, and drug efficacy. Wikidata currently contains 1502 items corresponding to human genetic variants,

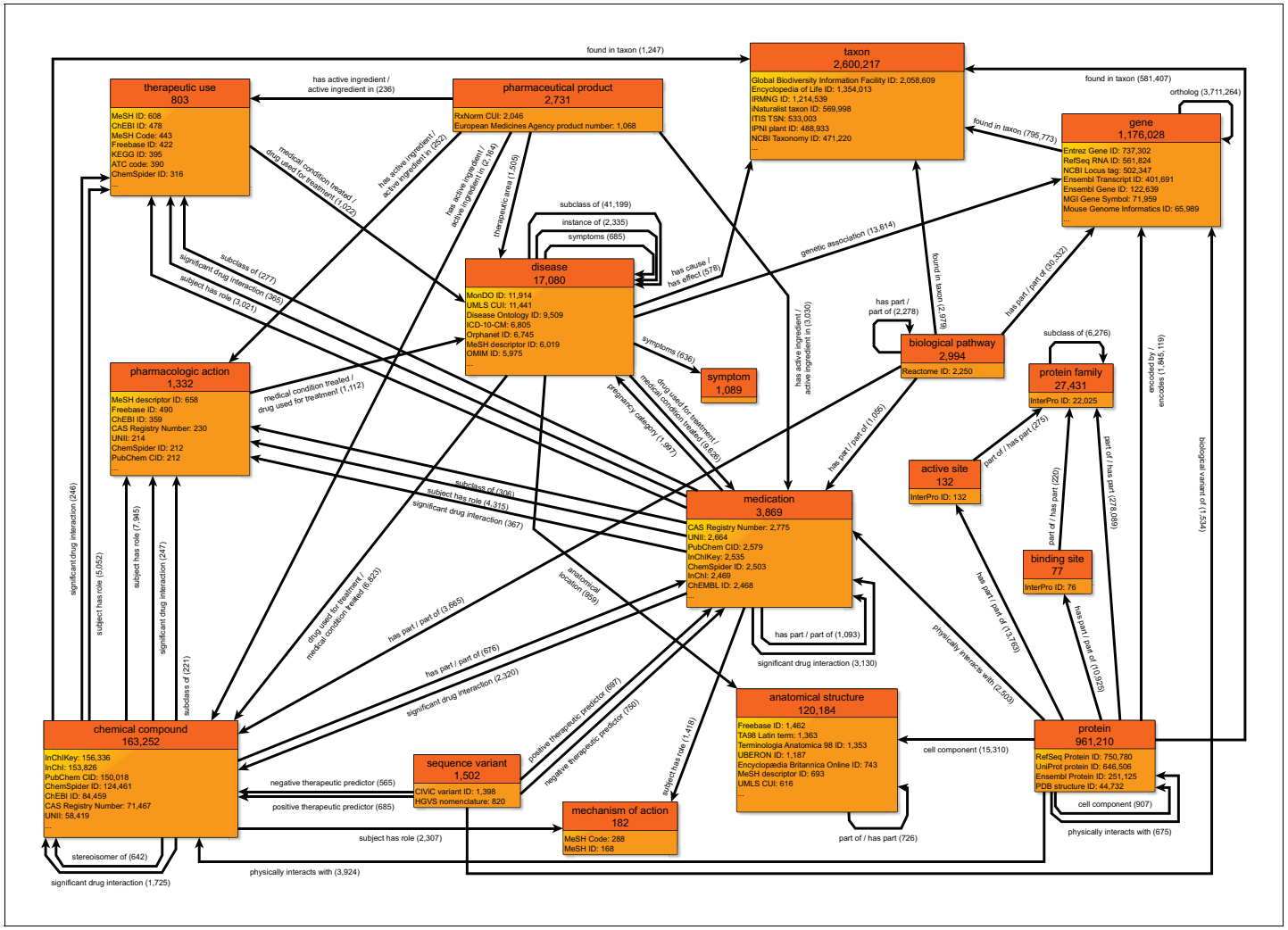

Figure 1. A simplified class-level diagram of the Wikidata knowledge graph for biomedical entities. Each box represents one type of biomedical entity. The header displays the name of that entity type (e.g., pharmaceutical product) and the number of Wikidata items for that entity type. The lower portion of each box displays a partial listing of attributes about each entity type and the number of Wikidata items for each attribute. Edges between boxes represent the number of Wikidata statements corresponding to each combination of subject type, predicate, and object type. For example, there are 1505 statements with 'pharmaceutical product' as the subject type, 'therapeutic area' as the predicate, and 'disease' as the object type. For clarity, edges for reciprocal relationships (e.g., 'has part' and 'part of') are combined into a single edge, and scientific articles (which are widely cited in statement references) have been omitted. All counts of Wikidata items are current as of September 2019. The most common data sources cited as references are available in ***Figure 1—source data 1***. Data are generated using the code in https://github.com/SuLab/genewikiworld (archived at ***Mayers et al., 2020***). A more complete version of this graph diagram can be found at https://commons.wikimedia.org/wiki/File:Biomedical_Knowledge_Graph_in_Wikidata.svg.

The online version of this article includes the following source data and figure supplement(s) for figure 1:

Source data 1. Most frequent data sources cited as references for the biomedical subset of the Wikidata knowledge graph shown in ***Figure 1***.
Figure supplement 1. Trends in Wikidata edits.

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 (***Wang et al., 2009***), RxNorm (***Nelson et al., 2011***), the IUPHAR Guide to Pharmacology (***Harding et al., 2018***; ***Pawson et al., 2014***; ***Southan et al., 2016***), NDF-RT (National Drug File – Reference Terminology), and LIPID MAPS (***Sud et al., 2007***). These items typically contain statements describing chemical structure and key physicochemical properties, and links to

databases with experimental data, such as MassBank (**Horai et al., 2010**; **Wohlgemuth et al., 2016**) and PDB Ligand (**Shin, 2004**), and toxicological information, such as the EPA CompTox Dashboard (**Williams et al., 2017**). 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 (**Fabregat et al., 2018**) and WikiPathways (**Slenter et al., 2018**). 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; **Sprague, 2017**), 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 (**Schriml et al., 2019**), with additional disease terms added from the Monarch Disease Ontology (**Mungall et al., 2017**). 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; **Ayers et al., 2019**), or both. Wikidata contains over 20 million items corresponding to publications across many domain areas, including a heavy emphasis on biomedical journal articles.

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. We then coded each bot to retrieve, transform, normalize and upload data from a primary resource to Wikidata 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 (https://github.com/SuLab/WikidataIntegrator; archived at **Burgstaller-Muehlbacher et al., 2020**). 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 of Wikidata

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, nor integrated in a reusable fashion.

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

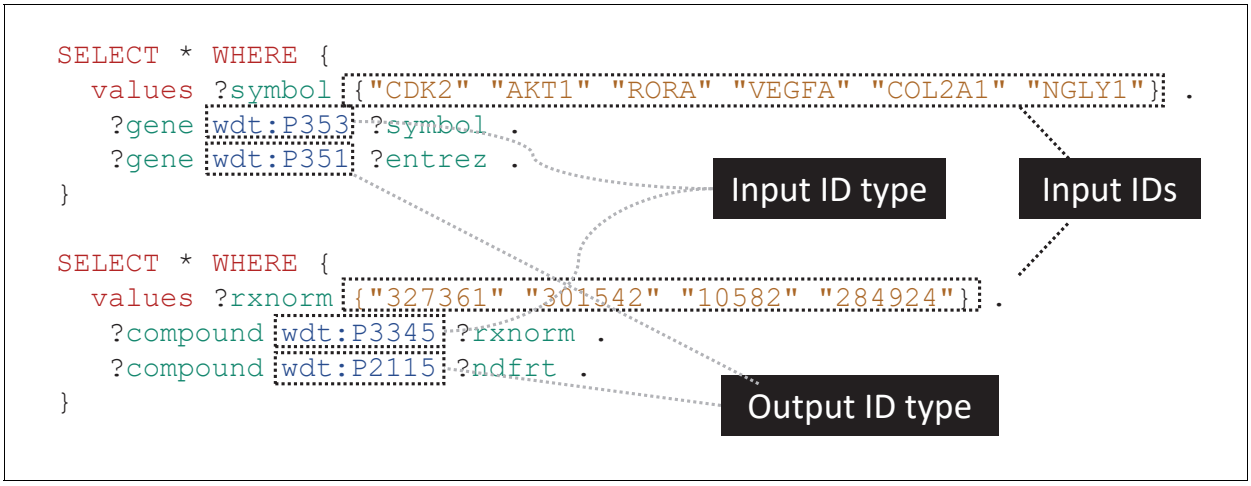

Figure 2. Generalizable SPARQL template for identifier translation. SPARQL is the primary query language for accessing Wikidata content. These simple SPARQL examples show how identifiers of any biological type can easily be translated using SPARQL queries. The top query demonstrates the translation of a small list of gene symbols (wdt:P353) to Entrez Gene IDs (wdt:P351), while the bottom example shows conversion of RxNorm concept IDs (wdt:P3345) to NDF-RT IDs (wdt:P2115). These queries can be submitted to the Wikidata Query Service (WDQS; https://query.wikidata.org/) to get real-time results. Translation to and from a wide variety of identifier types can be performed using slight modifications on these templates, and relatively simple extensions of these queries can filter mappings based on the statement references and/or qualifiers. A full list of Wikidata properties can be found at https://www.wikidata.org/wiki/Special:ListProperties. 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.

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 (***Xin et al., 2018***), BridgeDb (***van Iersel et al., 2010***), BioMart (***Smedley et al., 2015***), UMLS (***Bodenreider, 2004***), and NCI Thesaurus (***de Coronado et al., 2009***). 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. As a database of assertions and not of absolute truth, Wikidata is able to represent conflicting information (with provenance) when, for example, different curation authorities produce different mappings between entities. (However, as with any bioinformatics integration exercise, harmonization of cross-references between resources can include relationships other than 'exact match'. These instances can lead to Wikidata statements that are not explicitly declared, but rather the result of transitive inference.)

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***). They 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). They 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 they have collaborators with relevant expertise, they may specifically filter for proteins containing a serine-threonine kinase domain.

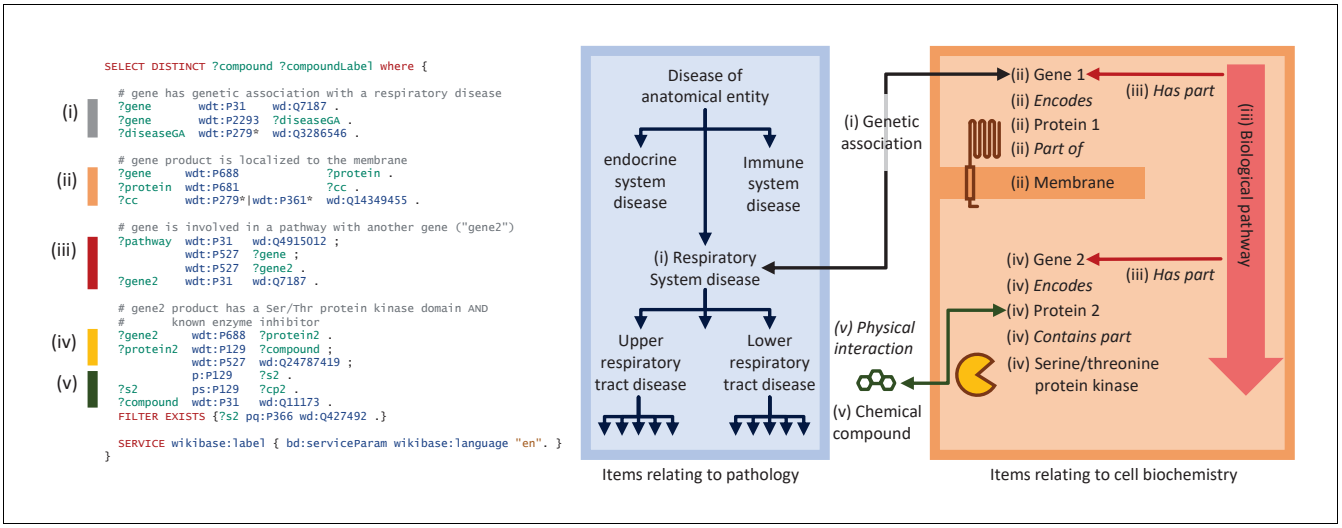

Figure 3. A representative SPARQL query that integrates data from multiple data resources and annotation types. This example integrative query incorporates data on genetic associations to disease, Gene Ontology annotations for cellular compartment, protein target information for compounds, pathway data, and protein domain information. Specifically, this query (depicted schematically at right) retrieves genes that are (i) associated with a respiratory system disease, (ii) that encode a membrane-bound protein, and (iii) that sit within the same biochemical pathway as (iv) a second gene encoding a protein with a serine-threonine kinase domain and (v) a known inhibitor, and reports a list of those inhibitors. Aspects related to Disease Ontology in blue; aspects related to biochemistry in red/orange; aspects related to chemistry in green. Properties are shown in italics. Real-time query results can be viewed at https://w.wiki/6pZ.

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 s 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. 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.

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 (see, for example, ***Bunt et al., 2012***; ***Gil et al., 2017***; ***Putman et al., 2019***; ***Putman et al., 2017***; ***Wang et al., 2016***). 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. Specifically, we designed a hybrid system that combines the aggregated community effort of many individuals with the reliability of expert curation. First, we created 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 (**Lewis et al., 2017**) and MeSH (https://www.nlm.nih.gov/mesh/meshhome.html). These cross references were primarily added by a small handful of users through a web interface focused on identifier mapping (**Manske, 2020**). Each cross reference was manually reviewed by DO expert curators, and 2007 of these mappings (98.9%) were deemed correct and therefore added to the ensuing DO release. 771 of the proposed mappings could not be easily validated using simple string matching, and 754 (97.8%) of these were ultimately accepted into DO. Each subsequent monthly report included a smaller number of added cross references to GARD and MeSH, as well as ORDO (**Maiella et al., 2018**), and OMIM (**Amberger and Hamosh, 2017**; **McKusick, 2007**), and these entries were incorporated after expert review at a high approval rate (>90%).

Addition of identifier mappings represents the most common community contribution, and likely the most accessible crowdsourcing task. However, Wikidata users also suggested numerous refinements to the ontology structure, including changes to the subclass relationships and the addition of new disease terms. These structural changes were more nuanced and therefore rarely incorporated into DO releases with no modifications. Nevertheless, 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 at https://github.com/SuLab/scheduled-bots (archived at **Stupp et al., 2020**) 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, curators, tools, and services.

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 (**Nielsen et al., 2017**). 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 (**Scholia, 2019**). 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 (https://tools.wmflabs.org/admin/tool/pathway-viewer) 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, **Scholia (2019)**. 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 were 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 (**Köhler et al., 2009**). On the back end, the latest version of Phenomizer uses BOQA, an algorithm that uses ontological structure in a Bayesian network (**Bauer et al., 2012**). For phenotype-based disease diagnosis, BOQA takes as input a list of phenotypes (using the Human Phenotype Ontology [HPO; **Köhler et al., 2017**]) and an association file between phenotypes and diseases. BOQA then suggests disease diagnoses based on semantic similarity (**Köhler et al., 2009**). Here, we studied whether phenotype-disease associations from

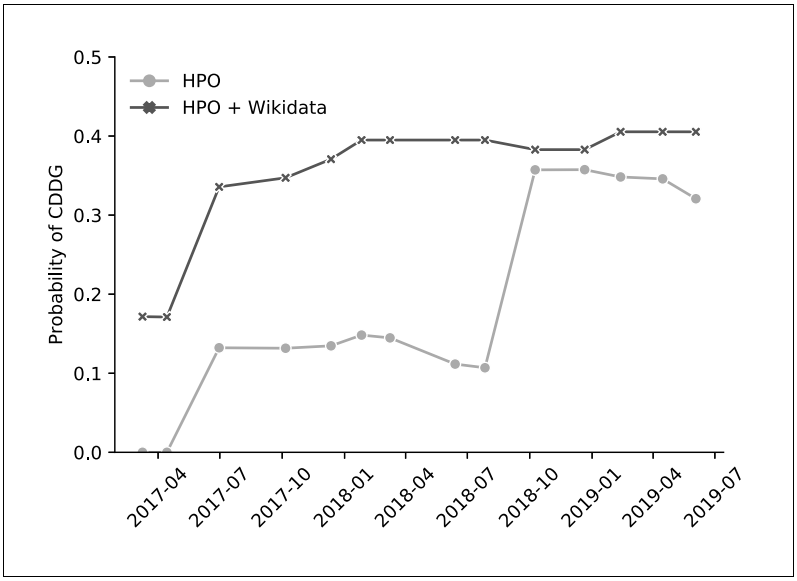

Figure 4. BOQA analysis of suspected cases of the disease Congenital Disorder of Deglycosylation (CDDG). We used an algorithm called BOQA to rank potential diagnoses based on clinical phenotypes. Here, clinical phenotypes from two cases of suspected CDDG patients were extracted from a published case report (***Caglayan et al., 2015***). These phenotypes were run through BOQA using phenotype-disease annotations from the Human Phenotype Ontology (HPO) alone, or from a combination of HPO and Wikidata. This analysis was tested using several versions of disease-phenotype annotations (shown along the x-axis). The probability score for CDDG is reported on the y-axis. These results demonstrate that the inclusion of Wikidata-based disease-phenotype annotations would have significantly improved the diagnosis predictions from BOQA at earlier time points prior to their official inclusion in the HPO annotation file. Details of this analysis can be found at https://github.com/SuLab/Wikidata-phenomizer (archived at ***Tu et al., 2020***).

Wikidata could improve BOQA's ability to make differential diagnoses for certain sets of phenotypes. We modified the BOQA codebase to accept arbitrary inputs and to be able to run from the command line (code available at https://github.com/SuLab/boqa; archived at ***Köhler and Stupp, 2020***) and also wrote a script to extract and incorporate the phenotype-disease annotations in Wikidata (code available at https://github.com/SuLab/Wikidata-phenomizer; archived at ***Tu et al., 2020***).

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) based on two papers describing patient phenotypes (***Enns et al., 2014***; ***Lam et al., 2017***). To see if the Wikidata-sourced annotations improved the ability of BOQA to diagnose CDDG, we ran our modified version using the phenotypes taken from a third publication describing two siblings with suspected cases of CDDG (***Caglayan et al., 2015***). Using these phenotypes and the annotation file supplemented with Wikidata-derived associations, BOQA 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. This example was specifically chosen to illustrate a favorable case, and the benefit of Wikidata would likely not currently generalize to a random sampling of other diseases. Nevertheless, we believe that this proof-of-concept demonstrates the value of the crowd-based Wikidata model and may motivate further community contributions.

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 (***Himmelstein et al., 2017***). 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 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 5***). We also performed this analysis using an external test set based on Drug Central, which showed a similar improvement in

Rephetio results over time (***Figure 5—figure supplement 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.

Outlook

We believe that the design of Wikidata is well-aligned with the FAIR data principles.

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 is listed in the life science registries FAIRsharing (https://fairsharing.org/; ***Sansone et al., 2019***) and Identifiers.org (***Wimalaratne et al., 2018***). 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 provides database dumps at least weekly (https://www.wikidata.org/wiki/Wikidata:Database_download), ensuring the long-term accessibility of the Wikidata knowledge graph independent of the organization and web application. Finally, Wikidata is also natively multilingual.

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 (***Jacobsen et al., 2018***). 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.

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 (see, for example, ***Reiser et al., 2016***). 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, albeit at the cost of not being able to incorporate ontologies or datasets with more restrictive licensing.

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 (***Stupp and Su, 2018***). 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; ***Das et al., 2017***; ***Lin et al., 2018***; ***Xiong et al., 2017***).

For those who subscribe to this vision for cultivating a FAIR and open graph of biomedical

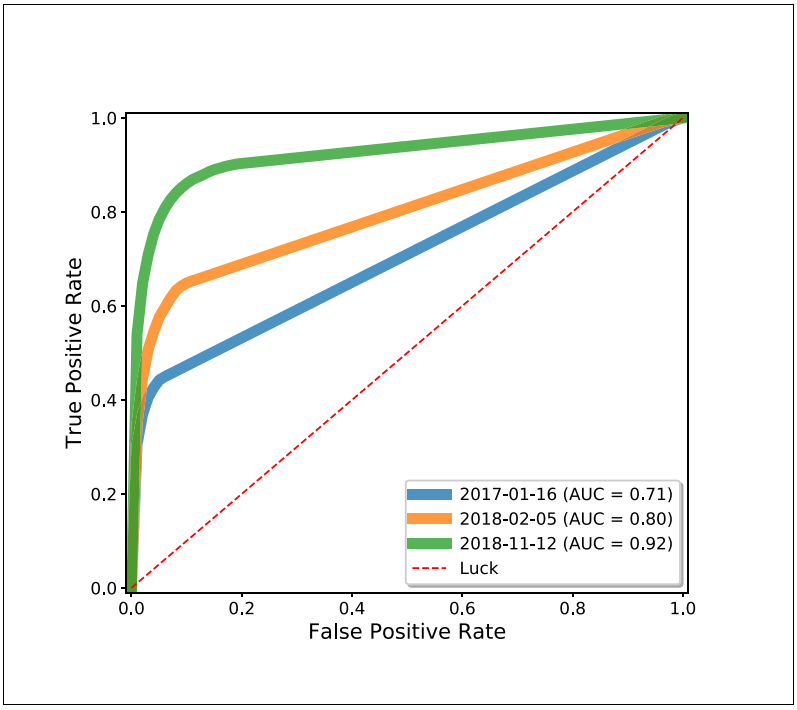

Figure 5. Drug repurposing using the Wikidata knowledge graph. We analyzed three snapshots of Wikidata using Rephetio, a graph-based algorithm for predicting drug repurposing candidates (***Himmelstein et al., 2017***). We evaluated the performance of the Rephetio algorithm on three historical versions 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 (archived at ***Mayers and Su, 2020***).

The online version of this article includes the following figure supplement(s) for figure 5:

Figure supplement 1. Drug repurposing using the Wikidata knowledge graph, evaluated using an external test set.

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 "move[s] the attribution from the legal realm into the social or ethical realm by pairing a permissive license with a strong moral entreaty' (***Cohen, 2013***). 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 , including the Disease Ontology (***Schriml et al., 2019***), Pfam (***El-Gebali et al., 2019***), Bgee (***Bastian et al., 2008***), WikiPathways (***Slenter et al., 2018***), Reactome (***Fabregat et al., 2018***), ECO (***Chibucos et al., 2014***), and CIViC (***Griffith et al., 2017***).

Second, informaticians can contribute to Wikidata by adding the results of data parsing and integration efforts to Wikidata as, for example, new Wikidata items, statements, or references. 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 performed 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. Indeed, the growth of biomedical data in Wikidata is driven not by any centralized or coordinated process, but rather the aggregated effort and priorities of Wikidata contributors themselves.

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.

Acknowledgements

The authors thank the thousands of Wikidata contributors for curating knowledge, both directly related and unrelated to this work, much of which has been organized under the WikiProjects for Molecular Biology, Chemistry and Medicine. The authors also thank the Wikimedia Foundation for financially supporting Wikidata, and many developers and administrators for maintaining Wikidata as a community resource.

Andra Waagmeester is at Micelio, Antwerp, Belgium
https://orcid.org/0000-0001-9773-4008

Gregory Stupp is in the Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, United States
https://orcid.org/0000-0002-0644-7212

Sebastian Burgstaller-Muehlbacher is in the Center for Integrative Bioinformatics Vienna, Max Perutz

Laboratories, University of Vienna and Medical University of Vienna, Vienna, Austria
https://orcid.org/0000-0003-4640-3510

Benjamin M Good is in the Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, United States
https://orcid.org/0000-0002-7334-7852

Malachi Griffith is in the McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, United States
https://orcid.org/0000-0002-6388-446X

Obi L Griffith is in the McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, United States
https://orcid.org/0000-0002-0843-4271

Kristina Hanspers is in the Institute of Data Science and Biotechnology, Gladstone Institutes, San Francisco, CA, United States
https://orcid.org/0000-0001-5410-599X

Henning Hermjakob is at the European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom
https://orcid.org/0000-0001-8479-0262

Toby S Hudson is in the School of Chemistry, University of Sydney, Sydney, Australia
https://orcid.org/0000-0002-3348-3622

Kevin Hybiske is in the Division of Allergy and Infectious Diseases, Department of Medicine, University of Washington, Seattle, WA, United States
https://orcid.org/0000-0002-2967-3079

Sarah M Keating is at the European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom
https://orcid.org/0000-0002-3356-3542

Magnus Manske is at the Wellcome Trust Sanger Institute, Hinxton, United Kingdom
https://orcid.org/0000-0001-5916-0947

Michael Mayers is in the Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, United States
https://orcid.org/0000-0002-7792-0150

Daniel Mietchen is in the School of Data Science, University of Virginia, Charlottesville, VA, United States
https://orcid.org/0000-0001-9488-1870

Elvira Mitraka is in the University of Maryland School of Medicine, Baltimore, MD, United States
https://orcid.org/0000-0003-0719-3485

Alexander R Pico is in the Institute of Data Science and Biotechnology, Gladstone Institutes, San Francisco, CA, United States
https://orcid.org/0000-0001-5706-2163

Timothy Putman is in the Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, United States
https://orcid.org/0000-0002-4291-0737

Anders Riutta is in the Institute of Data Science and Biotechnology, Gladstone Institutes, San Francisco, CA, United States
https://orcid.org/0000-0002-4693-0591

Núria Queralt-Rosinach is in the Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, United States
https://orcid.org/0000-0003-0169-8159

Lynn M Schriml is in the University of Maryland School of Medicine, Baltimore, MD, United States
https://orcid.org/0000-0001-8910-9851

Thomas Shafee is in the Department of Animal Plant and Soil Sciences, La Trobe University, Melbourne, Australia
https://orcid.org/0000-0002-2298-7593

Denise Slenter is in the Department of Bioinformatics-BiGCaT, NUTRIM, Maastricht University, Maastricht, Netherlands
https://orcid.org/0000-0001-8449-1318

Ralf Stephan is a retired researcher based in Berlin, Germany
https://orcid.org/0000-0002-4650-631X

Katherine Thornton is at Yale University Library, Yale University, New Haven, CT, United States
https://orcid.org/0000-0002-4499-0451

Ginger Tsueng is in the Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, United States
https://orcid.org/0000-0001-9536-9115

Roger Tu is in the Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, United States
https://orcid.org/0000-0002-7899-1604

Sabah Ul-Hasan is in the Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, United States
https://orcid.org/0000-0001-6334-452X

Egon Willighagen is in the Department of Bioinformatics-BiGCaT, NUTRIM, Maastricht University, Maastricht, Netherlands
https://orcid.org/0000-0001-7542-0286

Chunlei Wu is in the Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, United States
https://orcid.org/0000-0002-2629-6124

Andrew I Su is in the Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, United States
asu@scripps.edu
https://orcid.org/0000-0002-9859-4104

Author contributions: Andra Waagmeester, Conceptualization, Data curation, Software, Formal analysis, Validation, Visualization, Writing - review and editing; Gregory Stupp, Conceptualization, Data curation, Software, Formal analysis, Validation, Visualization, Writing - original draft; Sebastian Burgstaller-Muehlbacher, Conceptualization, Data curation, Software; Benjamin M Good, Conceptualization, Data curation, Software, Supervision, Project administration; Malachi Griffith, Kevin Hybiske, Data curation, Funding acquisition; Obi L Griffith, Data curation, Funding acquisition, Writing - review and editing; Kristina Hanspers, Sarah M

Keating, Magnus Manske, Timothy Putman, Anders Riutta, Nuria Queralt-Rosinach, Denise Slenter, Ginger Tsueng, Sabah Ul-Hasan, Egon Willighagen, Data curation, Software; Henning Hermjakob, Lynn M Schriml, Data curation, Supervision, Funding acquisition; Toby S Hudson, Elvira Mitraka, Ralf Stephan, Data curation, Validation; Michael Mayers, Data curation, Software, Formal analysis, Visualization; Daniel Mietchen, Data curation, Validation, Writing - review and editing; Alexander R Pico, Data curation, Supervision, Writing - original draft, Writing - review and editing; Thomas Shafee, Data curation, Visualization, Writing - review and editing; Katherine Thornton, Software, Validation; Roger Tu, Software, Formal analysis, Visualization; Chunlei Wu, Data curation, Software, Supervision, Funding acquisition; Andrew I Su, Conceptualization, Formal analysis, Supervision, Funding acquisition, Validation, Writing - original draft, Project administration, Writing - review and editing

Competing interests: The authors declare that no competing interests exist.

Received 09 October 2019
Accepted 28 February 2020
Published 17 March 2020

Funding

Funder	Grant reference number	Author
National Institute of General Medical Sciences	R01 GM089820	Andrew I Su
National Institute of General Medical Sciences	U54 GM114833	Henning Hermjakob Andrew I Su
National Institute of General Medical Sciences	R01 GM100039	Alexander R Pico
National Human Genome Research Institute	R00HG007940	Malachi Griffith
National Cancer Institute	U24CA237719	Malachi Griffith
V Foundation for Cancer Research	V2018-007	Malachi Griffith
National Institute of Allergy and Infectious Diseases	R01 AI126785	Kevin Hybiske
National Center for Advancing Translational Sciences	UL1 TR002550	Andrew I Su

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Decision letter and Author response
Decision letter https://doi.org/10.7554/eLife.52614.sa1
Author response https://doi.org/10.7554/eLife.52614.sa2

Additional files

Data availability

Links to all data and code used in this manuscript have been provided.

References

Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, Bolton E, Bourexis D, Brister JR, Bryant SH, Canese K, Cavanaugh M, Charowhas C, Clark K, Dondoshansky I, Feolo M, Fitzpatrick L, Funk K, Geer LY, Gorelenkov V, Graeff A, et al. 2018. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **46**:D8–D13. DOI: https://doi.org/10.1093/nar/gkx1095, PMID: 29140470

Amberger JS, Hamosh A. 2017. Searching Online Mendelian Inheritance in Man (OMIM): A knowledgebase of human genes and genetic phenotypes. *Current Protocols in Bioinformatics* **58**:27. DOI: https://doi.org/10.1002/cpbi.27

Ayers P, Mietchen D, Orlowitz J, Proffitt M, Rodlund S, Seiver E, Taraborelli D, Vershbow B. 2019. *WikiCite 2018-2019: Citations for the Sum of All Human Knowledge*: Wikimedia Foundation.

Bastian F, Parmentier G, Roux J, Moretti S, Laudet V, Robinson-Rechavi M. 2008. Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. In: Bairoch A, Cohen-Boulakia S, Froidevaux C (Eds). *Data Integration in the Life Sciences, Lecture Notes in Computer Science*. Berlin Heidelberg: Springer. p. 124–131 . DOI: https://doi.org/10.1007/978-3-540-69828-9_12

Bauer S, Köhler S, Schulz MH, Robinson PN. 2012. Bayesian ontology querying for accurate and noise-tolerant semantic searches. *Bioinformatics* **28**:2502–2508. DOI: https://doi.org/10.1093/bioinformatics/bts471, PMID: 22843981

Bodenreider O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research* **32**:267D–270. DOI: https://doi.org/10.1093/nar/gkh061

Bunt SM, Grumbling GB, Field HI, Marygold SJ, Brown NH, Millburn GH, FlyBase Consortium. 2012. Directly e-mailing authors of newly published papers encourages community curation. *Database* **2012**: bas024. DOI: https://doi.org/10.1093/database/bas024, PMID: 22554788

Burgstaller-Muehlbacher S, Waagmeester A, Mitraka E, Turner J, Putman T, Leong J, Naik C, Pavlidis P, Schriml L, Good BM, Su AI. 2016. Wikidata as a semantic framework for the Gene Wiki initiative. *Database* **2016**:baw015. DOI: https://doi.org/10.1093/database/baw015, PMID: 26989148

Burgstaller-Muehlbacher S, Waagmeester A, Stupp GS, Arrow T, konstin, jleong-ndn, Putman TE, Mystou, Mietchen D, Su AI, pdehaye, Leturia I, Vrandečić D, Andrea. 2020. SuLab/WikidataIntegrator 0.5.1. *Zenodo*. http://doi.org/10.5281/zenodo.3621065

Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Costanzo LD, Christie C, Duarte JM, Dutta S, Feng Z, Ghosh S, Goodsell DS, Green RK, Guranovic V, Guzenko D, Hudson BP, Liang Y, Lowe R, Peisach E, Periskova I, et al. 2019. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research* **47**:D520–D528. DOI: https://doi.org/10.1093/nar/gky949, PMID: 30357364

Caglayan AO, Comu S, Baranoski JF, Parman Y, Kaymakçalan H, Akgumus GT, Caglar C, Dolen D, Erson-Omay EZ, Harmanci AS, Mishra-Gorur K, Freeze HH, Yasuno K, Bilguvar K, Gunel M. 2015. NGLY1 mutation causes neuromotor impairment, intellectual disability, and neuropathy. *European Journal of Medical Genetics* **58**:39–43. DOI: https://doi.org/10.1016/j.ejmg.2014.08.008, PMID: 25220016

Chandras C, Weaver T, Zouberakis M, Smedley D, Schughart K, Rosenthal N, Hancock JM, Kollias G, Schofield PN, Aidinis V. 2009. Models for financial sustainability of biological databases and resources. *Database* **2009**:bap017. DOI: https://doi.org/10.1093/database/bap017, PMID: 20157490

Chibucos MC, Mungall CJ, Balakrishnan R, Christie KR, Huntley RP, White O, Blake JA, Lewis SE, Giglio M. 2014. Standardized description of scientific evidence using the Evidence Ontology (ECO). *Database* **2014**: bau075. DOI: https://doi.org/10.1093/database/bau075, PMID: 25052702

Cohen D. 2013. CC0 (+BY). https://dancohen.org/2013/11/26/cc0-by/

Das R, Dhuliawala S, Zaheer M, Vilnis L, Durugkar I, Krishnamurthy A, Smola A, McCallum A. 2017. Go for a walk and arrive at the answer: reasoning over paths in knowledge bases using reinforcement learning. *arXiv*. https://arxiv.org/abs/1711.05851.

de Coronado S, Wright LW, Fragoso G, Haber MW, Hahn-Dantona EA, Hartel FW, Quan SL, Safran T, Thomas N, Whiteman L. 2009. The NCI Thesaurus quality assurance life cycle. *Journal of Biomedical Informatics* **42**:530–539. DOI: https://doi.org/10.1016/j.jbi.2009.01.003, PMID: 19475726

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. *Nucleic Acids Research* **47**:D427–D432. DOI: https://doi.org/10.1093/nar/gky995, PMID: 30357350

Enns GM, Shashi V, Bainbridge M, Gambello MJ, Zahir FR, Bast T, Crimian R, Schoch K, Platt J, Cox R, Bernstein JA, Scavina M, Walter RS, Bibb A, Jones M, Hegde M, Graham BH, Need AC, Oviedo A, Schaaf CP, et al. 2014. Mutations in NGLY1 cause an inherited disorder of the endoplasmic reticulum-associated degradation pathway. *Genetics in Medicine* **16**:751–758. DOI: https://doi.org/10.1038/gim.2014.22, PMID: 24651605

Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, Milacic M, Roca CD, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Viteri G, Weiser J, Wu G, et al. 2018. The Reactome Pathway Knowledgebase. *Nucleic Acids Research* **46**:D649–D655. DOI: https://doi.org/10.1093/nar/gkx1132, PMID: 29145629

Gabella C, Durinx C, Appel R. 2018. Funding knowledgebases: towards a sustainable funding model for the UniProt use case. *F1000Research* **6**:2051. DOI: https://doi.org/10.12688/f1000research.12989.2

Gil Y, Garijo D, Ratnakar V, Khider D, Emile-Geay J, McKay N. 2017. A Controlled Crowdsourcing Approach for Practical Ontology Extensions and Metadata Annotations. In: d'Amato C, Fernandez M, Tamma V, Lecue F, Cudré-Mauroux P, Sequeda J, Lange C, Heflin J (Eds). *The Semantic Web – ISWC 2017, Lecture Notes in Computer Science*. Springer International Publishing. p. 231–246. DOI: https://doi.org/10.1007/978-3-319-68204-4

Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, Ainscough BJ, Ramirez CA, Rieke DT, Kujan L, Barnell EK, Wagner AH, Skidmore ZL, Wollam A, Liu CJ, Jones MR, Bilski RL, Lesurf R, Feng YY, Shah NM, et al. 2017. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature Genetics* **49**:170–174. DOI: https://doi.org/10.1038/ng.3774, PMID: 28138153

Harding SD, Sharman JL, Faccenda E, Southan C, Pawson AJ, Ireland S, Gray AJG, Bruce L, Alexander SPH, Anderton S, Bryant C, Davenport AP, Doerig C, Fabbro D, Levi-Schaffer F, Spedding M, Davies JA, NC-IUPHAR. 2018. The IUPHAR/BPS guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY. *Nucleic Acids Research* **46**:D1091–D1106. DOI: https://doi.org/10.1093/nar/gkx1121, PMID: 29149325

Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, Green A, Khankhanian P, Baranzini SE. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* **6**: e26726. DOI: https://doi.org/10.7554/eLife.26726, PMID: 28936969

Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, Ojima Y, Tanaka K, Tanaka S, Aoshima K, Oda Y, Kakazu Y, Kusano M, Tohge T, Matsuda F, Sawada Y, Hirai MY, Nakanishi H, Ikeda K, Akimoto N, et al. 2010. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry* **45**:703–714. DOI: https://doi.org/10.1002/jms.1777, PMID: 20623627

Jacobsen A, Kaliyaperumal R, Stupp GS, Schriml LM, Thompson M, Su AI, Roos M. 2018. Wikidata as an intuitive resource towards semantic data modeling in data FAIRification. In: *Proceedings of the 11th International Conference Semantic Web Applications and Tools for Life Sciences, {SWAT4LS} 2018, Antwerp, Belgium, December 3-6, 2018*. **2275** CEUR-WS.org.

Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, Mundlos C, Horn D, Mundlos S, Robinson PN. 2009. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics* **85**:457–464. DOI: https://doi.org/10.1016/j.ajhg.2009.09.003, PMID: 19800049

Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, Baynam G, Bello SM, Boerkoel CF, Boycott KM, Brudno M, Buske OJ, Chinnery PF, Cipriani V, Connell LE, Dawkins HJ, DeMare LE, Devereau AD, de Vries BB, Firth HV, et al. 2017. The Human Phenotype Ontology in 2017. *Nucleic Acids Research* **45**:D865–D876. DOI: https://doi.org/10.1093/nar/gkw1039, PMID: 27899602

Köhler S, Stupp GS. 2020. SuLab/boqa: release v1.0 2020-01-21. *Zenodo*. http://doi.org/10.5281/zenodo.3620979
Lam C, Ferreira C, Krasnewich D, Toro C, Latham L, Zein WM, Lehky T, Brewer C, Baker EH, Thurm A, Farmer CA, Rosenzweig SD, Lyons JJ, Schreiber JM, Gropman A, Lingala S, Ghany MG, Solomon B, Macnamara E, Davids M, et al. 2017. Prospective phenotyping of NGLY1-CDDG, the first congenital disorder of deglycosylation. *Genetics in Medicine* **19**: 160–168. DOI: https://doi.org/10.1038/gim.2016.75, PMID: 27388694
Lewis J, Snyder M, Hyatt-Knorr H. 2017. Marking 15 years of the Genetic and Rare Diseases Information Center. *Translational Science of Rare Diseases* **2**:77–88. DOI: https://doi.org/10.3233/TRD-170011, PMID: 29152459
Lin XV, Socher R, Xiong C. 2018. Multi-hop knowledge graph reasoning with reward shaping. *arXiv*. https://arxiv.org/abs/1808.10568.
Maiella S, Olry A, Hanauer M, Lanneau V, Lourghi H, Donadille B, Rodwell C, Köhler S, Seelow D, Jupp S, Parkinson H, Groza T, Brudno M, Robinson PN, Rath A. 2018. Harmonising phenomics information for a better interoperability in the rare disease field. *European Journal of Medical Genetics* **61**:706–714. DOI: https://doi.org/10.1016/j.ejmg.2018.01.013, PMID: 29425702
Manske M. 2020. Mix'n'match. https://tools.wmflabs.org/mix-n-match/#/
Mayers MD, Su AI. 2020. SuLab/WD-rephetio-analysis: v1.1 release on 2020-01-21. *Zenodo*. http://doi.org/10.5281/zenodo.3620811
Mayers MD, Stupp GS, Su AI. 2020. SuLab/genewikiworld: v1.1 release on 2020-01-21. *Zenodo*. http://doi.org/10.5281/zenodo.3620812
McKusick VA. 2007. Mendelian Inheritance In Man and its online version, OMIM. *The American Journal of Human Genetics* **80**:588–604. DOI: https://doi.org/10.1086/514346, PMID: 17357067
Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang HY, El-Gebali S, Fraser MI, Gough J, Haft DR, Huang H, Letunic I, Lopez R, Luciani A, Madeira F, Marchler-Bauer A, Mi H, Natale DA, et al. 2019. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research* **47**:D351–D360. DOI: https://doi.org/10.1093/nar/gky1100, PMID: 30398656
Mora-Cantallops M, Sánchez-Alonso S, García-Barriocanal E. 2019. A systematic literature review on Wikidata. *Data Technologies and Applications* **53**:250–268. DOI: https://doi.org/10.1108/DTA-12-2018-0110
Mungall CJ, McMurry JA, Köhler S, Balhoff JP, Borromeo C, Brush M, Carbon S, Conlin T, Dunn N, Engelstad M, Foster E, Gourdine JP, Jacobsen JO, Keith D, Laraway B, Lewis SE, NguyenXuan J, Shefchek K, Vasilevsky N, Yuan Z, et al. 2017. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research* **45**:D712–D722. DOI: https://doi.org/10.1093/nar/gkw1128, PMID: 27899636
Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. 2011. Normalized names for clinical drugs: RxNorm at 6 years. *Journal of the American Medical Informatics Association* **18**:441–448. DOI: https://doi.org/10.1136/amiajnl-2011-000116, PMID: 21515544
Nielsen FA, Mietchen D, Willighagen E. 2017. Scholia, Scientometrics and Wikidata. In: Blomqvist E, Hose K, Paulheim H, Ławrynowicz A, Ciravegna F, Hartig O (Eds). *The Semantic Web: ESWC 2017 Satellite Events, Lecture Notes in Computer Science*. Springer International Publishing. p. 237–259. DOI: https://doi.org/10.1007/978-3-319-70407-4_36
Pawson AJ, Sharman JL, Benson HE, Faccenda E, Alexander SP, Buneman OP, Davenport AP, McGrath JC, Peters JA, Southan C, Spedding M, Yu W, Harmar AJ, NC-IUPHAR. 2014. The IUPHAR/BPS guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands. *Nucleic Acids Research* **42**:D1098–D1106. DOI: https://doi.org/10.1093/nar/gkt1143, PMID: 24234439
Putman TE, Lelong S, Burgstaller-Muehlbacher S, Waagmeester A, Diesh C, Dunn N, Munoz-Torres M, Stupp GS, Wu C, Su AI, Good BM. 2017. WikiGenomes: an open web application for community consumption and curation of gene annotation data in Wikidata. *Database* **2017**:bax025. DOI: https://doi.org/10.1093/database/bax025
Putman T, Hybiske K, Jow D, Afrasiabi C, Lelong S, Cano MA, Wu C, Su AI. 2019. ChlamBase: a curated model organism database for the *Chlamydia* research community. *Database* **2019**:baz041. DOI: https://doi.org/10.1093/database/baz041, PMID: 30985891
Reiser L, Berardini TZ, Li D, Muller R, Strait EM, Li Q, Mezheritsky Y, Vetushko A, Huala E. 2016. Sustainable funding for biocuration: The Arabidopsis Information Resource (TAIR) as a case study of a subscription-based funding model. *Database* **2016**:baw018. DOI: https://doi.org/10.1093/database/baw018, PMID: 26989150
Sansone SA, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL, Thurston M, FAIRsharing Community. 2019. FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology* **37**:358–367. DOI: https://doi.org/10.1038/s41587-019-0080-8, PMID: 30940948
Scholia. 2019. Scholia, ACE Inhibitor Pathway. https://tools.wmflabs.org/scholia/pathway/Q29892242
Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, Felix V, Jeng L, Bearer C, Lichenstein R, Bisordi K, Campion N, Hyman B, Kurland D, Oates CP, Kibbey S, Sreekumar P, Le C, Giglio M, Greene C. 2019. Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Research* **47**:D955–D962. DOI: https://doi.org/10.1093/nar/gky1032, PMID: 30407550
Shin J-M. 2004. PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures. *Nucleic Acids Research* **33**:D238–D241. DOI: https://doi.org/10.1093/nar/gki059
Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Mélius J, Cirillo E, Coort SL, Digles D, Ehrhart F, Giesbertz P, Kalafati M, Martens M, Miller R, Nishida K, Rieswijk L, Waagmeester A, Eijssen LMT, Evelo CT, et al. 2018. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research* **46**:D661–D667. DOI: https://doi.org/10.1093/nar/gkx1064, PMID: 29136241
Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, Arnaiz O, Awedh MH, Baldock R, Barbiera G, Bardou P, Beck T, Blake A, Bonierbale M, Brookes AJ,

Bucci G, Buetti I, Burge S, Cabau C, Carlson JW, et al. 2015. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research* **43**:W589–W598. DOI: https://doi.org/10.1093/nar/gkv350, PMID: 25897122

Southan C, Sharman JL, Benson HE, Faccenda E, Pawson AJ, Alexander SP, Buneman OP, Davenport AP, McGrath JC, Peters JA, Spedding M, Catterall WA, Fabbro D, Davies JA, NC-IUPHAR. 2016. The IUPHAR/BPS guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. *Nucleic Acids Research* **44**:D1054–D1068. DOI: https://doi.org/10.1093/nar/gkv1037, PMID: 26464438

Sprague ER. 2017. ORCID. *Journal of the Medical Library Association* **105**:207–208. DOI: https://doi.org/10.5195/jmla.2017.89

Stupp GS, Su AI. 2018. Drug indications extracted from FAERS. *Zenodo*. http://doi.org/10.5281/zenodo.3620977

Stupp GS, Waagmeester A, Tsueng G, Pico AR, Tu R, Ul-Hasan S, Burgstaller-Muehlbacher S, Riutta A, Jacobson M, Su AI. 2020. SuLab/scheduled-bots: release v1.0 2020-01-21. *Zenodo*. http://doi.org/10.5281/zenodo.1436000

Sud M, Fahy E, Cotter D, Brown A, Dennis EA, Glass CK, Merrill AH, Murphy RC, Raetz CR, Russell DW, Subramaniam S. 2007. LMSD: LIPID MAPS Structure Database. *Nucleic Acids Research* **35**:D527–D532. DOI: https://doi.org/10.1093/nar/gkl838, PMID: 17098933

Tu R, Stupp GS, Su AI. 2020. SuLab/Wikidata-phenomizer: release v1.0 on 2020-01-15. *Zenodo*. http://doi.org/10.5281/zenodo.3609142

Turki H, Shafee T, Hadj Taieb MA, Ben Aouicha M, Vrandečić D, Das D, Hamdi H. 2019. Wikidata: a large-scale collaborative ontological medical database. *Journal of Biomedical Informatics* **99**:103292. DOI: https://doi.org/10.1016/j.jbi.2019.103292, PMID: 31557529

UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **47**:D506–D515. DOI: https://doi.org/10.1093/nar/gky1049, PMID: 30395287

Ursu O, Holmes J, Knockel J, Bologa CG, Yang JJ, Mathias SL, Nelson SJ, Oprea TI. 2017. DrugCentral: online drug compendium. *Nucleic Acids Research* **45**:D932–D939. DOI: https://doi.org/10.1093/nar/gkw993, PMID: 27789690

van Iersel MP, Pico AR, Kelder T, Gao J, Ho I, Hanspers K, Conklin BR, Evelo CT. 2010. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics* **11**:5. DOI: https://doi.org/10.1186/1471-2105-11-5, PMID: 20047655

Vrandečić D. 2012. Wikidata: a new platform for collaborative data collection. Proceedings of the 21st International Conference on World Wide Web. DOI: https://doi.org/10.1145/2187980.2188242

Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. 2009. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research* **37**:W623–W633. DOI: https://doi.org/10.1093/nar/gkp456, PMID: 19498078

Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu WT, Crüsemann M, Boudreau PD, Esquenazi E, Sandoval-Calderón M, Kersten RD, et al. 2016. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature Biotechnology* **34**:828–837. DOI: https://doi.org/10.1038/nbt.3597, PMID: 27504778

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* **3**:160018. DOI: https://doi.org/10.1038/sdata.2016.18, PMID: 26978244

Wilkinson MD, Dumontier M, Sansone SA, Bonino da Silva Santos LO, Prieto M, Batista D, McQuilton P, Kuhn T, Rocca-Serra P, Crosas M, Schultes E. 2019. Evaluating FAIR maturity through a scalable, automated, community-governed framework. *Scientific Data* **6**:1–12. DOI: https://doi.org/10.1038/s41597-019-0184-5, PMID: 31541130

Williams AJ, Grulke CM, Edwards J, McEachran AD, Mansouri K, Baker NC, Patlewicz G, Shah I, Wambaugh JF, Judson RS, Richard AM. 2017. The CompTox chemistry dashboard: a community data resource for environmental chemistry. *Journal of Cheminformatics* **9**:61. DOI: https://doi.org/10.1186/s13321-017-0247-6, PMID: 29185060

Willighagen E, Slenter D, Mietchen D, Evelo C, Nielsen F. 2018. Wikidata and scholia as a hub linking chemical knowledge. *Figshare*. https://figshare.com/articles/Wikidata_and_Scholia_as_a_hub_linking_chemical_knowledge/6356027.

Wimalaratne SM, Juty N, Kunze J, Janée G, McMurry JA, Beard N, Jimenez R, Grethe JS, Hermjakob H, Martone ME, Clark T. 2018. Uniform resolution of compact identifiers for biomedical data. *Scientific Data* **5**:180029. DOI: https://doi.org/10.1038/sdata.2018.29, PMID: 29737976

Wohlgemuth G, Mehta SS, Mejia RF, Neumann S, Pedrosa D, Pluskal T, Schymanski EL, Willighagen EL, Wilson M, Wishart DS, Arita M, Dorrestein PC, Bandeira N, Wang M, Schulze T, Salek RM, Steinbeck C, Nainala VC, Mistrik R, Nishioka T, et al. 2016. SPLASH, a hashed identifier for mass spectra. *Nature Biotechnology* **34**:1099–1101. DOI: https://doi.org/10.1038/nbt.3689, PMID: 27824832

Xin J, Afrasiabi C, Lelong S, Adesara J, Tsueng G, Su AI, Wu C. 2018. Cross-linking BioThings APIs through JSON-LD to facilitate knowledge exploration. *BMC Bioinformatics* **19**:30. DOI: https://doi.org/10.1186/s12859-018-2041-5, PMID: 29390967

Xiong W, Hoang T, Wang WY. 2017. DeepPath: a reinforcement learning method for knowledge graph reasoning. *arXiv*. https://arxiv.org/abs/1707.06690.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, Gil L, Gordon L, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, To JK, Laird MR, et al. 2018. Ensembl 2018. *Nucleic Acids Research* **46**:D754–D761. DOI: https://doi.org/10.1093/nar/gkx1098, PMID: 29155950