Wikidata as a FAIR knowledge graph for the life sciences

Andra Waagmeester[1,*], Gregory Stupp[2,*], Sebastian Burgstaller-Muehlbacher[2,¶], Benjamin M. Good[2], Malachi Griffith[3], Obi Griffith[3], Kristina Hanspers[4], Henning Hermjakob[5], Kevin Hybiske[6], Sarah M. Keating[5], Magnus Manske[7], Michael Mayers[2], Elvira Mitraka[8], Alexander R. Pico[4], Timothy Putman[2], Anders Riutta[4], Núria Queralt-Rosinach[2], Lynn M. Schriml[8], Denise Slenter[9], Ginger Tsueng[2], Roger Tu[2], Egon Willighagen[9], Chunlei Wu[2], Andrew I. Su[2,§]

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

1

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.

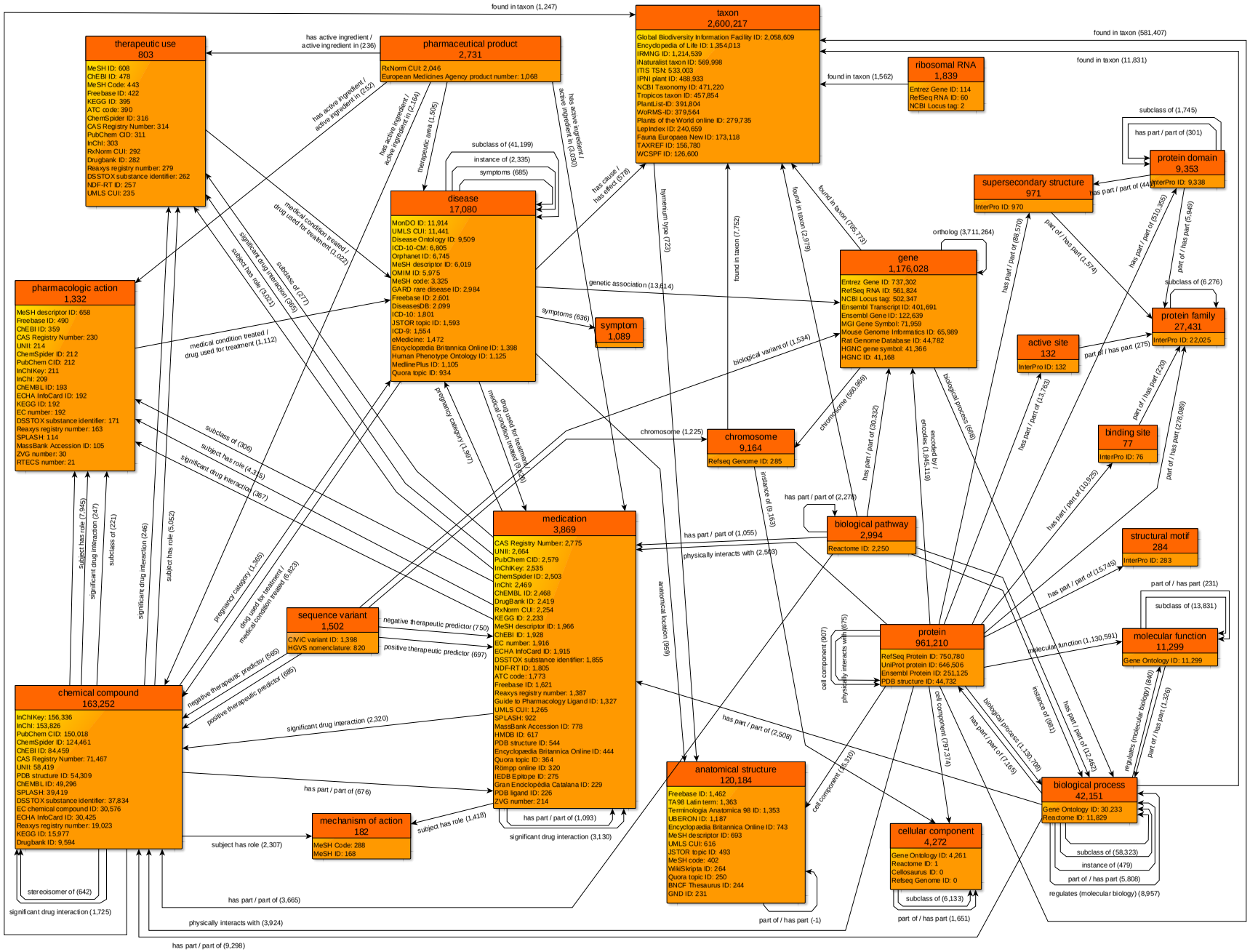

taxon
2,600,217
Global Biodiversity Information Facility ID: 2,058,609
Encyclopedia of Life ID: 1,354,013
IRMNG ID: 1,214,539
iNaturalist taxon ID: 569,998
ITIS TSN: 533,003
IPNI plant ID: 488,933
NCBI Taxonomy ID: 471,220
Tropicos taxon ID: 457,854
PlantList-ID: 391,804
WoRMS-ID: 379,564
Plants of the World online ID: 279,735
LepIndex ID: 240,659
Fauna Europaea New ID: 173,118
TAXREF ID: 156,780
WCSPF ID: 126,600
therapeutic use
803
MeSH ID: 608
ChEBI ID: 478
MeSH Code: 443
Freebase ID: 422
KEGG ID: 395
ATC code: 390
ChemSpider ID: 316
CAS Registry Number: 314
PubChem CID: 311
InChI: 303
RxNorm CUI: 292
Drugbank ID: 282
Reaxys registry number: 279
DSSTOX substance identifier: 262
NDF-RT ID: 257
UMLS CUI: 235
pharmaceutical product
2,731
RxNorm CUI: 2,046
European Medicines Agency product number: 1,068
ribosomal RNA
1,839
Entrez Gene ID: 114
RefSeq RNA ID: 60
NCBI Locus tag: 2
disease
17,080
MonDO ID: 11,914
UMLS CUI: 11,441
Disease Ontology ID: 9,509
ICD-10-CM: 6,805
Orphanet ID: 6,745
MeSH descriptor ID: 6,019
OMIM ID: 5,975
MeSH code: 3,325
GARD rare disease ID: 2,984
Freebase ID: 2,601
DiseasesDB: 2,099
ICD-10: 1,801
JSTOR topic ID: 1,593
ICD-9: 1,554
eMedicine: 1,472
Encyclopædia Britannica Online ID: 1,398
Human Phenotype Ontology ID: 1,125
MedlinePlus ID: 1,105
Quora topic ID: 934
gene
1,176,028
Entrez Gene ID: 737,302
RefSeq RNA ID: 561,824
NCBI Locus tag: 502,347
Ensembl Transcript ID: 401,691
Ensembl Gene ID: 122,639
MGI Gene Symbol: 71,959
Mouse Genome Informatics ID: 65,989
Rat Genome Database ID: 44,782
HGNC gene symbol: 41,366
HGNC ID: 41,168
supersecondary structure
971
InterPro ID: 970
protein domain
9,353
InterPro ID: 9,338
protein family
27,431
InterPro ID: 22,025
active site
132
InterPro ID: 132
binding site
77
InterPro ID: 76
structural motif
284
InterPro ID: 283
pharmacologic action
1,332
MeSH descriptor ID: 658
Freebase ID: 490
ChEBI ID: 359
CAS Registry Number: 230
UNII: 214
ChemSpider ID: 212
PubChem CID: 212
InChIKey: 211
InChI: 209
ChEMBL ID: 193
ECHA InfoCard ID: 192
KEGG ID: 192
EC number: 192
DSSTOX substance identifier: 171
Reaxys registry number: 163
SPLASH: 114
MassBank Accession ID: 105
ZVG number: 30
RTECS number: 21
symptom
1,089
chromosome
9,164
Refseq Genome ID: 285
biological pathway
2,994
Reactome ID: 2,250
medication
3,869
CAS Registry Number: 2,775
UNII: 2,664
PubChem CID: 2,579
InChIKey: 2,535
ChemSpider ID: 2,503
InChI: 2,469
ChEMBL ID: 2,468
DrugBank ID: 2,419
RxNorm CUI: 2,254
KEGG ID: 2,233
MeSH descriptor ID: 1,966
ChEBI ID: 1,928
EC number: 1,916
ECHA InfoCard ID: 1,915
DSSTOX substance identifier: 1,855
NDF-RT ID: 1,805
ATC code: 1,773
Freebase ID: 1,621
Reaxys registry number: 1,387
Guide to Pharmacology Ligand ID: 1,327
UMLS CUI: 1,265
SPLASH: 922
MassBank Accession ID: 778
HMDB ID: 617
PDB structure ID: 544
Encyclopædia Britannica Online ID: 444
Quora topic ID: 364
Römpp online ID: 320
IEDB Epitope ID: 275
Gran Enciclopèdia Catalana ID: 229
PDB ligand ID: 226
ZVG number: 214
sequence variant
1,502
CIVIC variant ID: 1,398
HGVS nomenclature: 820
protein
961,210
RefSeq Protein ID: 750,780
UniProt protein ID: 646,506
Ensembl Protein ID: 251,125
PDB structure ID: 44,732
molecular function
11,299
Gene Ontology ID: 11,299
chemical compound
163,252
InChIKey: 156,336
InChI: 153,826
PubChem CID: 150,018
ChemSpider ID: 124,461
ChEBI ID: 84,459
CAS Registry Number: 71,467
UNII: 58,419
PDB structure ID: 54,309
ChEMBL ID: 49,296
SPLASH: 39,419
DSSTOX substance identifier: 37,834
EC chemical compound ID: 30,576
ECHA InfoCard ID: 30,425
Reaxys registry number: 19,023
KEGG ID: 15,977
Drugbank ID: 9,594
anatomical structure
120,184
Freebase ID: 1,462
TA98 Latin term: 1,363
Terminologia Anatomica 98 ID: 1,353
UBERON ID: 1,187
Encyclopædia Britannica Online ID: 743
MeSH descriptor ID: 693
UMLS CUI: 616
JSTOR topic ID: 493
MeSH code: 402
WikiSkripta ID: 264
Quora topic ID: 250
BNCF Thesaurus ID: 244
GND ID: 231
biological process
42,151
Gene Ontology ID: 30,233
Reactome ID: 11,829
cellular component
4,272
Gene Ontology ID: 4,261
Reactome ID: 1
Cellosaurus ID: 0
Refseq Genome ID: 0
mechanism of action
182
MeSH Code: 288
MeSH ID: 168
found in taxon (1,247)
has active ingredient / active ingredient in (236)
found in taxon (581,407)
found in taxon (11,831)
found in taxon (1,562)
subclass of (1,745)
has part / part of (301)
subclass of (41,199)
instance of (2,335)
symptoms (685)
has active ingredient / active ingredient in (252)
has active ingredient / active ingredient in (2,264)
therapeutic area (1,595)
has active ingredient / active ingredient in (3,030)
has cause / has effect (578)
hymenum type (723)
found in taxon (1,752)
found in taxon (2,979)
found in taxon (795,773)
medical condition treated / drug used for treatment (2,022)
significant drug interaction (365)
subject has role (3,021)
subclass of (277)
genetic association (13,614)
symptoms (636)
orthologs (3,711,264)
has part / part of (88,570)
part of / has part (1,574)
has part / part of (510,265)
part of / has part (5,949)
subclass of (6,276)
medical condition treated / drug used for treatment (1,112)
biological variant of (1,534)
part of / has part (275)
part of / has part (220)
chromosome (1,225)
chromosome (550,969)
has part / part of (30,332)
encoded by / encodes (1,945,110)
biological process (668)
has part / part of (13,763)
part of / has part (278,088)
subclass of (306)
subject has role (4,335)
significant drug interaction (367)
pregnancy category (1,937)
drug used for treatment / medical condition treated (8,835)
instance of (9,153)
has part / part of (2,278)
has part / part of (10,925)
subject has role (7,945)
significant drug interaction (247)
subclass of (221)
significant drug interaction (246)
subject has role (6,052)
has part / part of (1,055)
physically interacts with (2,503)
has part / part of (15,745)
part of / has part (231)
subclass of (13,831)
pregnancy category (1,365)
drug used for treatment / medical condition treated (6,823)
negative therapeutic predictor (750)
positive therapeutic predictor (697)
anatomical location (959)
cell component (907)
physically interacts with (675)
molecular function (1,130,591)
negative therapeutic predictor (565)
positive therapeutic predictor (695)
significant drug interaction (2,320)
has part / part of (2,508)
instance of (981)
has part / part of (12,462)
regulates (molecular biology) (940)
part of / has part (1,326)
cell component (797,374)
biological process (1,130,788)
has part / part of (7,165)
cell component (15,310)
has part / part of (676)
subject has role (1,418)
has part / part of (1,093)
significant drug interaction (3,130)
subject has role (2,307)
subclass of (58,323)
instance of (479)
part of / has part (5,808)
regulates (molecular biology) (8,957)
stereoisomer of (642)
significant drug interaction (1,725)
has part / part of (3,665)
physically interacts with (3,924)
part of / has part (-1)
subclass of (6,133)
part of / has part (1,651)
has part / part of (9,298)

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.

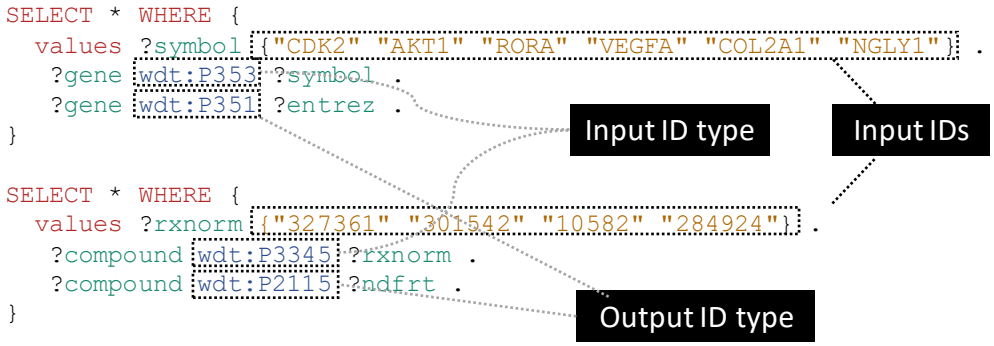


Figure 2. Generalizable SPARQL template for identifier translation. This simple example shows how identifiers of any biological type can easily be translated using SPARQL queries. These queries operate on Wikidata properties for gene symbols (wdt:P353) and Entrez Gene IDs (wdt:P351) (top), and RxNorm 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

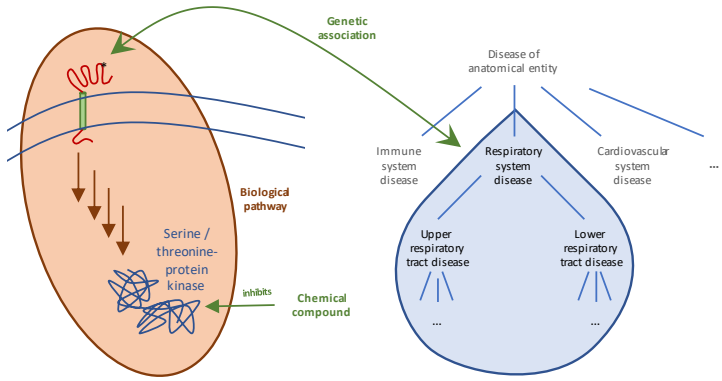
Wikidata contains a much broader set of information than just identifier cross-references. Having biomedical data in one centralized data resource facilitates powerful integrative queries that span multiple domain areas and data sources. Performing these integrative queries through Wikidata obviates 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. 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.




```
SELECT DISTINCT ?compound ?compoundLabel where {

  # gene has genetic association with a respiratory disease
  ?gene       wdt:P31    wd:Q7187 .
  ?gene       wdt:P2293  ?diseaseGA .
  ?diseaseGA  wdt:P279*  wd:Q3286546 .

  # gene product is localized to the membrane
  ?gene     wdt:P688             ?protein .
  ?protein  wdt:P681             ?cc .
  ?cc       wdt:P279*|wdt:P361*  wd:Q14349455 .

  # gene is involved in a pathway with another gene ("gene2")
  ?pathway  wdt:P31   wd:Q4915012 ;
            wdt:P527  ?gene ;
            wdt:P527  ?gene2 .
  ?gene2    wdt:P31   wd:Q7187 .

  # gene2 product has a Ser/Thr protein kinase domain AND known enzyme inhibitor
  ?gene2     wdt:P688  ?protein2 .
  ?protein2  wdt:P129  ?compound ;
             wdt:P527  wd:Q24787419 ;
             p:P129    ?s2 .
  ?s2        ps:P129   ?cp2 .
  ?compound  wdt:P31   wd:Q11173 .
  FILTER EXISTS {?s2 pq:P366 wd:Q427492 .}

  SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}
```

Figure 3. A representative SPARQL query that integrates data from multiple data resources and annotation types. 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.

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.

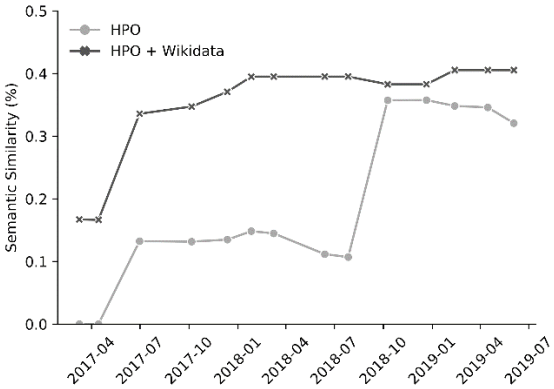


Figure 4. Phenomizer analysis of suspected cases of CDDG. Clinical phenotypes from two cases of suspected CDDG patients were extracted from a published case report [62]. These phenotypes were run through the Phenomizer tool using phenotype-disease annotations from HPO alone, or from a combination of 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.

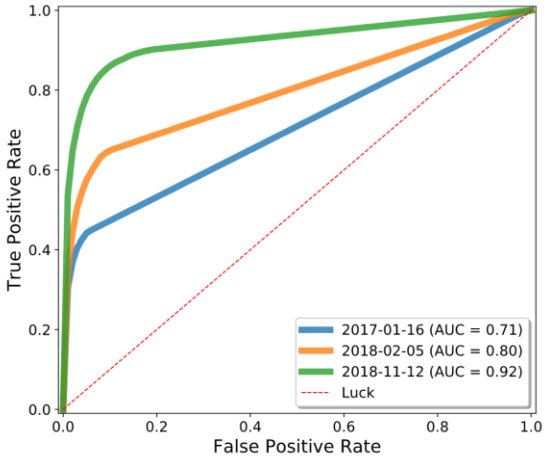
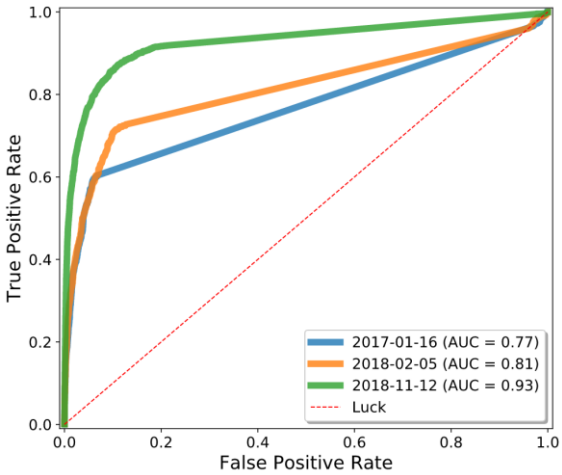


Figure 5. Drug repurposing using the Wikidata knowledge graph. We analyzed three snapshots of Wikidata using Rephetio, a graph-based algorithm for predicting drug repurposing candidates [63]. We evaluated the performance of the Rephetio algorithm on three historical versions of the Wikidata knowledge 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 Wikdiata. 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

1. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, 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, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018. PMCID: PMC4792175

2. Evaluating FAIR Maturity Through a Scalable, Automated, Community-Governed Framework | bioRxiv [Internet]. [cited 2019 Jul 31]. Available from: https://www.biorxiv.org/content/10.1101/649202v1

3. 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 JOB, Keith D, Laraway B, Lewis SE, NguyenXuan J, Shefchek K, Vasilevsky N, Yuan Z, Washington N, Hochheiser H, Groza T, Smedley D, Robinson PN, Haendel MA. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res. 2017 04;45(D1):D712–D722. PMCID: PMC5210586

4. Gabella C, Durinx C, Appel R. Funding knowledgebases: Towards a sustainable funding model for the UniProt use case. F1000Research [Internet]. 2018 Mar 22 [cited 2019 Aug 26];6. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5747334/ PMCID: PMC5747334

5. Chandras C, Weaver T, Zouberakis M, Smedley D, Schughart K, Rosenthal N, Hancock JM, Kollias G, Schofield PN, Aidinis V. Models for financial sustainability of biological databases and resources. Database [Internet]. 2009 Jan 1 [cited 2019 Aug 26];2009. Available from: https://academic.oup.com/database/article/doi/10.1093/database/bap017/357253

6. Vrandečić D. Wikidata: A New Platform for Collaborative Data Collection. Proc 21st Int Conf World Wide Web [Internet]. New York, NY, USA: ACM; 2012 [cited 2019 Aug 1]. p. 1063–1064. Available from: http://doi.acm.org/10.1145/2187980.2188242

7. Wikidata Statistics [Internet]. [cited 2019 Sep 11]. Available from: https://tools.wmflabs.org/wikidata-todo/stats.php

8. Mora-Cantallops M, Sánchez-Alonso S, García-Barriocanal E. A systematic literature review on Wikidata. Data Technol Appl [Internet]. 2019 Jul 1 [cited 2019 Sep 6]; Available from: https://www.emerald.com/insight/content/doi/10.1108/DTA-12-2018-0110/full/html

9. Wikidata Query Service [Internet]. [cited 2019 Jul 31]. Available from: https://query.wikidata.org/

10. Burgstaller-Muehlbacher S, Waagmeester A, Mitraka E, Turner J, Putman T, Leong J, Naik C, Pavlidis P, Schriml L, Good BM, Su AI. Wikidata as a semantic framework for the Gene Wiki initiative. Database J Biol Databases Curation. 2016;2016. PMCID: PMC4795929

11. Willighagen E, Slenter D, Mietchen D, Evelo C, Nielsen F. Wikidata and Scholia as a hub linking chemical knowledge [Internet]. 2018 [cited 2019 Aug 23]. Available from:

https://figshare.com/articles/Wikidata_and_Scholia_as_a_hub_linking_chemical_knowledge/6356027

12. Turki H, Shafee T, Taieb MAH, Aouicha MB, Vrandečić D, Das D, Hamdi H. Wikidata: A large-scale collaborative ontological medical database. J Biomed Inform. 2019 Sep 23;103292.

13. Putman TE, Lelong S, Burgstaller-Muehlbacher S, Waagmeester A, Diesh C, Dunn N, Munoz-Torres M, Stupp GS, Wu C, Su AI, Good BM. WikiGenomes: an open web application for community consumption and curation of gene annotation data in Wikidata. Database J Biol Databases Curation. 2017 01;2017(1). PMCID: PMC5467579

14. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2018 04;46(D1):D8–D13. PMCID: PMC5753372

15. 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, Lavidas I, Liu Z, Loveland JE, Maurel T, McLaren W, Moore B, Mudge J, Murphy DN, Newman V, Nuhn M, Ogeh D, Ong CK, Parker A, Patricio M, Riat HS, Schuilenburg H, Sheppard D, Sparrow H, Taylor K, Thormann A, Vullo A, Walts B, Zadissa A, Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Aken BL, Cunningham F, Yates A, Flicek P. Ensembl 2018. Nucleic Acids Res. 2018 04;46(D1):D754–D761. PMCID: PMC5753206

16. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019 Jan 8;47(D1):D506–D515.

17. Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang H-Y, 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, Necci M, Nuka G, Orengo C, Pandurangan AP, Paysan-Lafosse T, Pesseat S, Potter SC, Qureshi MA, Rawlings ND, Redaschi N, Richardson LJ, Rivoire C, Salazar GA, Sangrador-Vegas A, Sigrist CJA, Sillitoe I, Sutton GG, Thanki N, Thomas PD, Tosatto SCE, Yong S-Y, Finn RD. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res. 2019 Jan 8;47(D1):D351–D360.

18. 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, Randle C, Rose A, Sekharan M, Shao C, Tao Y-P, Valasatava Y, Voigt M, Westbrook J, Young J, Zardecki C, Zhuravleva M, Kurisu G, Nakamura H, Kengaku Y, Cho H, Sato J, Kim JY, Ikegawa Y, Nakagawa A, Yamashita R, Kudou T, Bekker G-J, Suzuki H, Iwata T, Yokochi M, Kobayashi N, Fujiwara T, Velankar S, Kleywegt GJ, Anyango S, Armstrong DR, Berrisford JM, Conroy MJ, Dana JM, Deshpande M, Gane P, Gáborová R, Gupta D, Gutmanas A, Koča J, Mak L, Mir S, Mukhopadhyay A, Nadzirin N, Nair S, Patwardhan A, Paysan-Lafosse T, Pravda L, Salih O, Sehnal D, Varadi M, Vařeková R, Markley JL, Hoch JC, Romero PR, Baskaran K, Maziuk D, Ulrich EL, Wedell JR, Yao H, Livny M, Ioannidis YE. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. 2019 Jan 8;47(D1):D520–D528.

19. 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 Y-Y, Shah NM, Bonakdar M, Trani L, Matlock M, Ramu A, Campbell KM, Spies GC, Graubert AP, Gangavarapu K, Eldred JM, Larson DE, Walker JR, Good BM, Wu C, Su

AI, Dienstmann R, Margolin AA, Tamborero D, Lopez-Bigas N, Jones SJM, Bose R, Spencer DH, Wartman LD, Wilson RK, Mardis ER, Griffith OL. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet. 2017 Jan 31;49(2):170–174. PMCID: PMC5367263

20. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res. 2009 Jul 1;37(suppl_2):W623–W633.

21. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. J Am Med Inform Assoc. 2011 Jul 1;18(4):441–448.

22. 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. The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY. Nucleic Acids Res. 2018 04;46(D1):D1091–D1106. PMCID: PMC5753190

23. Southan C, Sharman JL, Benson HE, Faccenda E, Pawson AJ, Alexander SPH, Buneman OP, Davenport AP, McGrath JC, Peters JA, Spedding M, Catterall WA, Fabbro D, Davies JA, NC-IUPHAR. The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. Nucleic Acids Res. 2016 Jan 4;44(D1):D1054-1068. PMCID: PMC4702778

24. Pawson AJ, Sharman JL, Benson HE, Faccenda E, Alexander SPH, Buneman OP, Davenport AP, McGrath JC, Peters JA, Southan C, Spedding M, Yu W, Harmar AJ, NC-IUPHAR. The IUPHAR/BPS Guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands. Nucleic Acids Res. 2014 Jan;42(Database issue):D1098-1106. PMCID: PMC3965070

25. UMLS Metathesaurus - NDFRT (National Drug File - Reference Terminology) - Synopsis [Internet]. [cited 2019 Sep 9]. Available from: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/index.html

26. Sud M, Fahy E, Cotter D, Brown A, Dennis EA, Glass CK, Merrill AH, Murphy RC, Raetz CRH, Russell DW, Subramaniam S. LMSD: LIPID MAPS structure database. Nucleic Acids Res. 2007 Jan 1;35(suppl_1):D527–D532.

27. 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, Maoka T, Takahashi H, Ara T, Sakurai N, Suzuki H, Shibata D, Neumann S, Iida T, Tanaka K, Funatsu K, Matsuura F, Soga T, Taguchi R, Saito K, Nishioka T. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom. 2010;45(7):703–714.

28. 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, Fiehn O. SPLASH, a hashed identifier for mass spectra. Nat Biotechnol. 2016 Nov 8;34:1099–1101.

29. Shin J-M, Cho D-H. PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures. Nucleic Acids Res. 2005 Jan 1;33(suppl_1):D238–D241.

30. Williams AJ, Grulke CM, Edwards J, McEachran AD, Mansouri K, Baker NC, Patlewicz G, Shah I, Wambaugh JF, Judson RS, Richard AM. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. J Cheminformatics. 2017 Nov 28;9(1):61.

31. 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, Stein L, Hermjakob H, D'Eustachio P. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018 Jan 4;46(D1):D649–D655. PMCID: PMC5753187

32. 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, Pico AR, Willighagen EL. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. 2018 Jan 4;46(D1):D661–D667. PMCID: PMC5753270

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

34. 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. Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Res. 2019 Jan 8;47(D1):D955–D962.

35. Ayers P, Mietchen D, Orlowitz J, Proffitt M, Rodlund S, Seiver E, Taraborelli D, Vershbow B. WikiCite 2018-2019: Citations for the sum of all human knowledge [Internet]. 2019 [cited 2019 Sep 6]. Available from: https://figshare.com/articles/WikiCite_2018-2019_Citations_for_the_sum_of_all_human_knowledge/8947451

36. Wikidata:WikiProject Molecular biology - Wikidata [Internet]. [cited 2019 Jul 29]. Available from: https://www.wikidata.org/wiki/Wikidata:WikiProject_Molecular_biology

37. A Wikidata Python module integrating the MediaWiki API and the Wikidata SPARQL endpoint: SuLab/WikidataIntegrator [Internet]. Su Lab; 2019 [cited 2019 Jul 23]. Available from: https://github.com/SuLab/WikidataIntegrator

38. Xin J, Afrasiabi C, Lelong S, Adesara J, Tsueng G, Su AI, Wu C. Cross-linking BioThings APIs through JSON-LD to facilitate knowledge exploration. BMC Bioinformatics. 2018 01;19(1):30. PMCID: PMC5796402

39. van Iersel MP, Pico AR, Kelder T, Gao J, Ho I, Hanspers K, Conklin BR, Evelo CT. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. BMC Bioinformatics. 2010 Jan 4;11(1):5.

40. 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, Chelala C, Chrysostomou C, Cittaro D, Collin O, Cordova R, Cutts RJ, Dassi E, Genova AD, Djari A, Esposito A, Estrella H, Eyras E, Fernandez-Banet J, Forbes S, Free RC, Fujisawa T, Gadaleta E, Garcia-Manteiga JM, Goodstein D, Gray K, Guerra-Assunção JA,

Haggarty B, Han D-J, Han BW, Harris T, Harshbarger J, Hastings RK, Hayes RD, Hoede C, Hu S, Hu Z-L, Hutchins L, Kan Z, Kawaji H, Keliet A, Kerhornou A, Kim S, Kinsella R, Klopp C, Kong L, Lawson D, Lazarevic D, Lee J-H, Letellier T, Li C-Y, Lio P, Liu C-J, Luo J, Maass A, Mariette J, Maurel T, Merella S, Mohamed AM, Moreews F, Nabihoudine I, Ndegwa N, Noirot C, Perez-Llamas C, Primig M, Quattrone A, Quesneville H, Rambaldi D, Reecy J, Riba M, Rosanoff S, Saddiq AA, Salas E, Sallou O, Shepherd R, Simon R, Sperling L, Spooner W, Staines DM, Steinbach D, Stone K, Stupka E, Teague JW, Dayem Ullah AZ, Wang J, Ware D, Wong-Erasmus M, Youens-Clark K, Zadissa A, Zhang S-J, Kasprzyk A. The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. 2015 Jul 1;43(W1):W589–W598.

41. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-270. PMCID: PMC308795

42. de Coronado S, Wright LW, Fragoso G, Haber MW, Hahn-Dantona EA, Hartel FW, Quan SL, Safran T, Thomas N, Whiteman L. The NCI Thesaurus quality assurance life cycle. J Biomed Inform. 2009 Jun;42(3):530–539. PMID: 19475726

43. List of Properties - Wikidata [Internet]. [cited 2019 Aug 23]. Available from: https://www.wikidata.org/wiki/Special:ListProperties

44. Gil Y, Garijo D, Ratnakar V, Khider D, Emile-Geay J, McKay N. 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, editors. Semantic Web – ISWC 2017. Springer International Publishing; 2017. p. 231–246.

45. Bunt SM, Grumbling GB, Field HI, Marygold SJ, Brown NH, Millburn GH, FlyBase Consortium. Directly e-mailing authors of newly published papers encourages community curation. Database J Biol Databases Curation. 2012;2012:bas024. PMCID: PMC3342516

46. 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 W-T, Crüsemann M, Boudreau PD, Esquenazi E, Sandoval-Calderón M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu C-C, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw C-C, Yang Y-L, Humpf H-U, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, Klitgaard A, Larson CB, Boya P CA, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams PG, Dai J, Neupane R, Gurr J, Rodríguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan BM, Almaliti J, Allard P-M, Phapale P, Nothias L-F, Alexandrov T, Litaudon M, Wolfender J-L, Kyle JE, Metz TO, Peryea T, Nguyen D-T, VanLeer D, Shinn P, Jadhav A, Müller R, Waters KM, Shi W, Liu X, Zhang L, Knight R, Jensen PR, Palsson BØ, Pogliano K, Linington RG, Gutiérrez M, Lopes NP, Gerwick WH, Moore BS, Dorrestein PC, Bandeira N. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol. 2016 Aug;34(8):828–837.

47. Putman T, Hybiske K, Jow D, Afrasiabi C, Lelong S, Cano MA, Stupp GS, Waagmeester A, Good BM, Wu C, Su AI. ChlamBase: a curated model organism database for the Chlamydia research community. Database J Biol Databases Curation. 2019 01;2019. PMCID: PMC6580685

48. Lewis J, Snyder M, Hyatt-Knorr H. Marking 15 years of the Genetic and Rare Diseases Information Center. Transl Sci Rare Dis. 2017 May 25;2(1–2):77–88. PMCID: PMC5685198

49. Medical Subject Headings - Home Page [Internet]. [cited 2019 Aug 27]. Available from: https://www.nlm.nih.gov/mesh/meshhome.html

50. 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. Harmonising phenomics information for a better interoperability in the rare disease field. Eur J Med Genet. 2018 Nov;61(11):706–714. PMID: 29425702

51. Amberger JS, Hamosh A. Searching Online Mendelian Inheritance in Man (OMIM): A Knowledgebase of Human Genes and Genetic Phenotypes. Curr Protoc Bioinforma. 2017 27;58:1.2.1-1.2.12. PMCID: PMC5662200

52. McKusick VA. Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet. 2007 Apr;80(4):588–604. PMCID: PMC1852721

53. GeneWiki Scheduled Bots. Contribute to SuLab/scheduled-bots development by creating an account on GitHub [Internet]. Su Lab; 2019 [cited 2019 Aug 23]. Available from: https://github.com/SuLab/scheduled-bots

54. fnielsen/scholia [Internet]. GitHub. [cited 2019 Sep 27]. Available from: https://github.com/fnielsen/scholia

55. Scholia [Internet]. [cited 2019 Oct 1]. Available from: https://tools.wmflabs.org/scholia/pathway/

56. Tool information: pathway-viewer - Wikimedia Toolforge [Internet]. [cited 2019 Sep 27]. Available from: https://tools.wmflabs.org/admin/tool/pathway-viewer

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

58. 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 HJS, DeMare LE, Devereau AD, de Vries BBA, Firth HV, Freson K, Greene D, Hamosh A, Helbig I, Hum C, Jähn JA, James R, Krause R, F. Laulederkind SJ, Lochmüller H, Lyon GJ, Ogishima S, Olry A, Ouwehand WH, Pontikos N, Rath A, Schaefer F, Scott RH, Segal M, Sergouniotis PI, Sever R, Smith CL, Straub V, Thompson R, Turner C, Turro E, Veltman MWM, Vulliamy T, Yu J, von Ziegenweidt J, Zankl A, Züchner S, Zemojtel T, Jacobsen JOB, Groza T, Smedley D, Mungall CJ, Haendel M, Robinson PN. The Human Phenotype Ontology in 2017. Nucleic Acids Res. 2017 Jan 4;45(D1):D865–D876.

59. Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, Mundlos C, Horn D, Mundlos S, Robinson PN. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet. 2009 Oct;85(4):457–464. PMCID: PMC2756558

60. 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

61. 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

62. 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. NGLY1 mutation causes neuromotor impairment, intellectual disability, and neuropathy. Eur J Med Genet. 2015 Jan;58(1):39–43. PMCID: PMC4804755

63. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, Green A, Khankhanian P, Baranzini SE. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife. 2017 22;6. PMCID: PMC5640425

64. Ursu O, Holmes J, Knockel J, Bologa CG, Yang JJ, Mathias SL, Nelson SJ, Oprea TI. DrugCentral: online drug compendium. Nucleic Acids Res. 2017 Jan 4;45(D1):D932–D939.

65. Union PO of the E. Turning FAIR into reality : final report and action plan from the European Commission expert group on FAIR data. [Internet]. 2018 [cited 2019 Aug 23]. Available from: https://publications.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en/format-PDF

66. Sansone S-A, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL, Thurston M. FAIRsharing as a community approach to standards, repositories and policies. Nat Biotechnol. 2019 Apr;37(4):358–367.

67. Wimalaratne SM, Juty N, Kunze J, Janée G, McMurry JA, Beard N, Jimenez R, Grethe JS, Hermjakob H, Martone ME, Clark T. Uniform resolution of compact identifiers for biomedical data. Sci Data. 2018 May 8;5:180029.

68. Wikidata:Database download - Wikidata [Internet]. [cited 2019 Aug 8]. Available from: https://www.wikidata.org/wiki/Wikidata:Database_download

69. Jacobsen A. Wikidata as an intuitive resource towards semantic data modeling in data FAIRification. 2018; Available from: http://ceur-ws.org/Vol-2275/short1.pdf

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

71. Reiser L, Berardini TZ, Li D, Muller R, Strait EM, Li Q, Mezheritsky Y, Vetushko A, Huala E. Sustainable funding for biocuration: The Arabidopsis Information Resource (TAIR) as a case study of a subscription-based funding model. Database J Biol Databases Curation. 2016;2016. PMCID: PMC4795935

72. Stupp GS, Su AI. Drug Indications Extracted from FAERS [Internet]. Zenodo; 2018 [cited 2019 Jun 27]. Available from: https://zenodo.org/record/1436000#.XRVY5-hKguU

73. Das R, Dhuliawala S, Zaheer M, Vilnis L, Durugkar I, Krishnamurthy A, Smola A, McCallum A. Go for a Walk and Arrive at the Answer: Reasoning Over Paths in Knowledge Bases using Reinforcement Learning. ArXiv171105851 Cs [Internet]. 2017 Nov 15 [cited 2019 May 6]; Available from: http://arxiv.org/abs/1711.05851

74. Xiong W, Hoang T, Wang WY. DeepPath: A Reinforcement Learning Method for Knowledge Graph Reasoning. ArXiv170706690 Cs [Internet]. 2017 Jul 20 [cited 2019 May 6]; Available from: http://arxiv.org/abs/1707.06690

75. Lin XV, Socher R, Xiong C. Multi-Hop Knowledge Graph Reasoning with Reward Shaping. ArXiv180810568 Cs [Internet]. 2018 Aug 30 [cited 2019 May 6]; Available from: http://arxiv.org/abs/1808.10568

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/

78. 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. The Pfam protein families database in 2019. Nucleic Acids Res. 2019 Jan 8;47(D1):D427–D432.

79. Bastian F, Parmentier G, Roux J, Moretti S, Laudet V, Robinson-Rechavi M. Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. In: Bairoch A, Cohen-Boulakia S, Froidevaux C, editors. Data Integr Life Sci. Springer Berlin Heidelberg; 2008. p. 124–131.

80. Chibucos MC, Mungall CJ, Balakrishnan R, Christie KR, Huntley RP, White O, Blake JA, Lewis SE, Giglio M. Standardized description of scientific evidence using the Evidence Ontology (ECO). Database [Internet]. 2014 Jan 1 [cited 2019 Aug 8];2014. Available from: https://academic.oup.com/database/article/doi/10.1093/database/bau075/2634798