Original article

Centralizing content and distributing labor: a community model for curating the very long tail of microbial genomes

Tim E. Putman[1], Sebastian Burgstaller-Muehlbacher[1], Andra Waagmeester[2], Chunlei Wu[1], Andrew I. Su[1] and Benjamin M. Good[1,*]

[1]Department of Molecular and Experimental Medicine, the Scripps Research Institute, La Jolla, USA, [2]Micelio, Antwerp, Belgium

*Corrsponding author: Tel: (858) 784-2079; Email: bgood@scripps.edu
Correspondence may also be addressed to Andrew I. Su. Tel: (858) 784-2079; Email: asu@scripps.edu

Abstract

The last 20 years of advancement in sequencing technologies have led to sequencing thousands of microbial genomes, creating mountains of genetic data. While efficiency in generating the data improves almost daily, applying meaningful relationships between taxonomic and genetic entities on this scale requires a structured and integrative approach. Currently, knowledge is distributed across a fragmented landscape of resources from government-funded institutions such as National Center for Biotechnology Information (NCBI) and UniProt to topic-focused databases like the ODB3 database of prokaryotic operons, to the supplemental table of a primary publication. A major drawback to large scale, expert-curated databases is the expense of maintaining and extending them over time. No entity apart from a major institution with stable long-term funding can consider this, and their scope is limited considering the magnitude of microbial data being generated daily. Wikidata is an openly editable, semantic web compatible framework for knowledge representation. It is a project of the Wikimedia Foundation and offers knowledge integration capabilities ideally suited to the challenge of representing the exploding body of information about microbial genomics. We are developing a microbial specific data model, based on Wikidata's semantic web compatibility, which represents bacterial species, strains and the gene and gene products that define them. Currently, we have loaded 43 694 gene and 37 966 protein items for 21 species of bacteria, including the human pathogenic bacteria *Chlamydia trachomatis*. Using this pathogen as an example, we explore complex interactions between the pathogen, its host, associated genes, other microbes, disease and drugs using the Wikidata SPARQL endpoint. In our next phase of development, we will add another 99 bacterial genomes and their gene and gene products, totaling ~900,000 additional entities. This aggregation of knowledge will be a

platform for community-driven collaboration, allowing the networking of microbial genetic data through the sharing of knowledge by both the data and domain expert.

Introduction

The relatively small and non-repetitive nature of microbial genomes, coupled with the rapid advancement of sequencing technology in the last decade, have led to the generation of a staggering amount of bacterial genome records. The National Center for Biotechnology Information (NCBI) Genome Database currently maintains genome records for over ~3000 high quality reference and representative genome assemblies and another ~50 000 incomplete assemblies. The existing collections of genomes are just the beginning; The Earth Microbiome Project (1) http://www.earthmicrobiome.org) is in the early stages of analyzing and cataloguing over ~200 000 environmental samples from around the world, and estimates that this will result in the sequencing of ~500 000 reconstructed microbial genomes (1). Making sense out of this abundance of data, while a daunting challenge, will generate a wealth of knowledge for the microbial and human genomic research community.

For microbial genomes, as well as most other biological data, knowledge is distributed across resources that occupy the full spectrum from very large, broad coverage, centralized, major government-funded institutions such as NCBI (www.NCBI.nlm.nih.gov/gene, (2)) and UniProt (www.uniprot.org, (3)) to boutique, topic-focused databases like the ODB3 database of prokaryotic operons (www.http://oper ondb.jp/, (4)), to the unstructured primary literature. The ability to process data smoothly from across that spectrum would greatly increase the efficiency of microbial research giving the researcher the ability to explore questions, based on existing evidence-based data, before any experimentation.

An example of such a question might be, 'what other microorganisms influence the persistence of an infection by a human pathogen such as *Chlamydia*, and by what mechanism?'. An expert may generate hypothetical answers to this question by blending their knowledge with information spread through the literature and various databases. As an example, the statements illustrated in Figure 1, originating from multiple sources, including primary literature (5–8), and structured databases such as NCBI Gene, UniProt, Drugbank (www.drugbank.ca/, (9)) and BRENDA (http://www.brenda-enzymes.info/, (10)), link together to yield the hypothesis that co-infection by *Prevotella spp., Clostridiales spp. and Escherichia coli* in the vaginal microbiome increase the persistence of infection through the generation of indole (6,11,12), a key substrate in the tryptophan biosynthesis pathway. Experts, like Caldwell and colleagues (6), have done the experiments and the literature research to generate the hypothesis that there is a greater risk of clearance failure, leading to persistent infection that should be treated appropriately, when these other indole-creating microbes are present (6,8).

By pulling these pieces of knowledge together into a common database, with defined connections between them, a list of the taxa involved can be generated as candidate answers to the above question with a single query. Once this is achieved and new data is added, the network grows and the collective benefit grows as well. The complicated and disordered is given order in a central container with a mechanism for sifting through it, giving the microbial researcher a powerful tool for making sense of the published data.

Model organism databases such as the Mouse Genome Database (http://nar.oxfordjournals.org/content/43/D1/D726.short) Mouse Genome Informatis (MGI) (http://www.informatics.jax.org), would greatly aid researcher's ability to unlock connections between microbes and the organisms they interact with. However, such large data warehouses are typically maintained by expensive teams of data and domain experts. The immense scale of microbial data is economically incompatible with this kind of centrally funded approach, and the same resolution would not be achieved. In one way or another, the greater scientific community, encompassing both active scientists and interested members of the general public must be empowered to contribute their mental energy in a community-wide collaborative effort (13). Here, we propose that Wikidata may provide the means to achieve this goal.

Wikidata is a new, centralized, yet openly editable platform for semantic knowledge representation that is maintained by the Wikimedia Foundation (the same entity that maintains all of the 200+ different language Wikipedias). Centralizing structured knowledge in this open database generates the opportunity to distribute the labor of data curation across a far broader community than was before realistic. In doing so, it offers a new approach to the knowledge integration problem that is ideally suited to the challenge of representing the exploding body of information about microbial genomics. Here, we describe the initial work of building a Wikidata-based representation of microbial genetics.
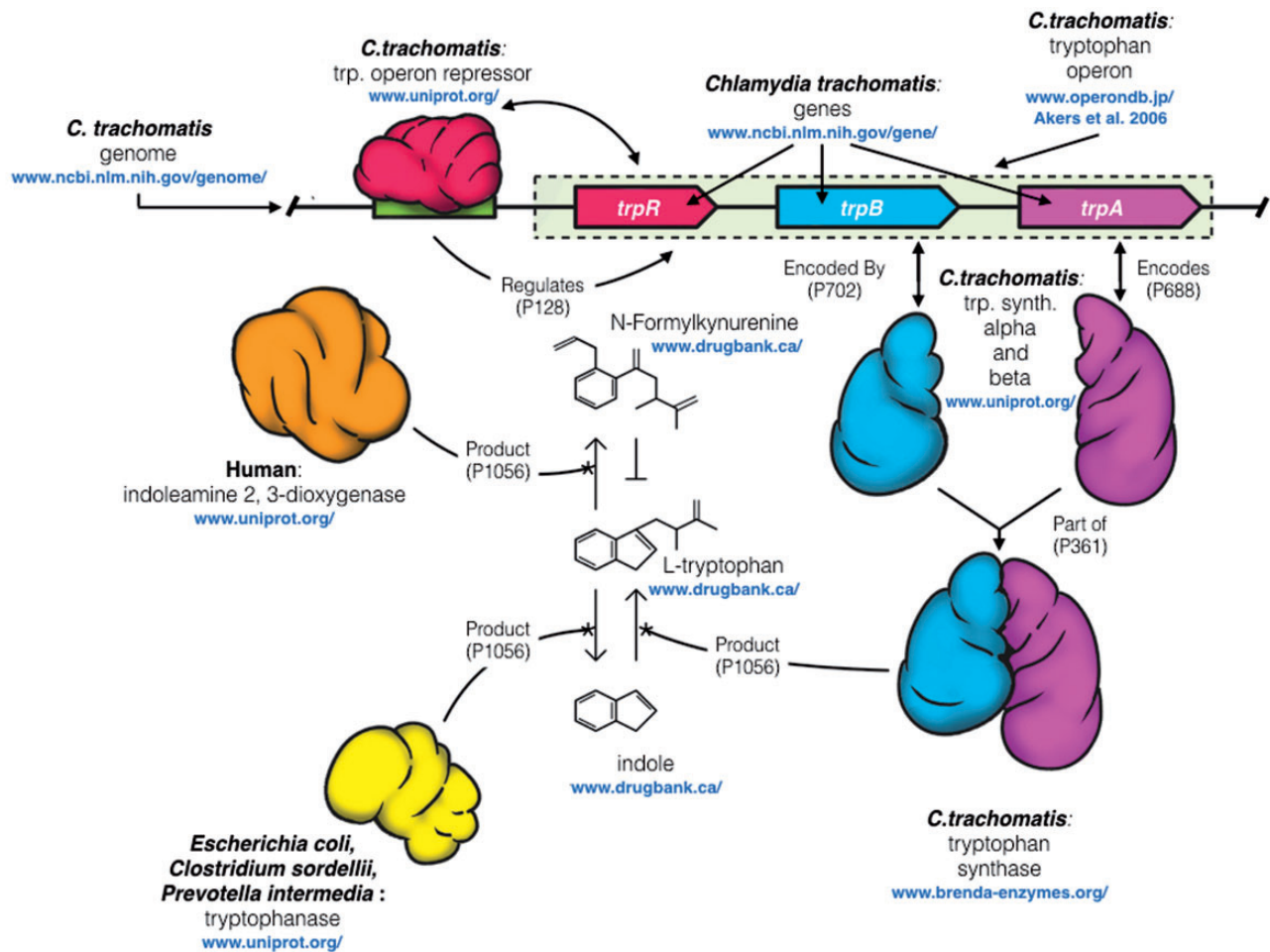
Figure 1. Illustration of the complex network of interacting entities between human, chlamydial, and other microbial species in the urogenital microbiome. When a human epithelial cell is infected by *C. trachomatis*, it responds by depleting the cell of L-tryptophan, an essential amino acid for chlamydial growth, through IFN-γ mediated expression of the tryptophan degrading enzyme indoleamine 2,3-dioxygenase (IDO)(orange)(5,6). IDO degrades tryptophan to *N*-Formylkynurenine, a tryptophan precursor that *C. trachomatis* is not capable of converting into tryptophan. Often this clears the infection, but episodically *C. trachomatis* rescues itself from this host defense by converting exogenous indole into L-tryptophan through gene expression regulated by its *trp operon*(8). Several experiments support the hypothesis that the likely source of exogenous indole is from other microbes in a perturbed vaginal microbiome; as part of L-tryptophan degradation via the pyruvate pathway. Microbes producing tryptophanase (yellow), an enzyme that degrades L-tryptophan to indole and pyruvate are commonly found in the urinary tract of patients also presenting with bacterial vaginosis (BV)(7). Examples of indole producers, commonly associated with BV in the female urogenital tract include *Prevotella spp., E. coli and Clostridiales spp.* (11,12, 23). Blue URLs indicate the various resources that maintain the data. The arrows between entities indicate the properties used to define their relationships once aggregated in Wikidata.

Wikidata as a centralized microbial database

A centralized resource for microbial genomics will need to capture a wide variety of different kinds of entities and relationships to support useful queries. Rather than attempt to build a system that models all of this complexity upfront, we are taking the approach of seeding the openly extensible Wikidata database with the beginnings of this model and thus encouraging the broader community to see the opportunity to collaborate on its evolution. Wikidata provides an ideal technical and social platform for undertaking this project. Its schema-free nature naturally supports data model changes and its open, wiki-based nature supports constructs such as 'talk pages', 'watchlists' and 'wikiprojects' that have proven effective in facilitating the attainment of community consensus over time in other open projects such as the GeneWiki (14, 15).

On a technical level, Wikidata is fundamentally a graph database (16). It has no overarching schema to ordain what kind of data elements can be linked together. In contrast to a standard relational model, this empowers database contributors to freely expand connections as they see fit. The only constraints on the structure of the graph are socially imposed. To add new content to Wikidata, would-be editors need to either use existing properties or make a request to the community that new properties be created. Property requests are openly discussed for a period

of time until a consensus has been reached on their value to the system and how they should be applied. This request process allows the community to control the emerging structure of the graph and reduce problems that might arise from completely unrestricted editing.

As a starting point for seeding the collaborative creation of a centralized microbial database in Wikidata, we established the structures needed to represent the entities and relations depicted in Figure 1. In the context of Wikidata, this work amounts to the creation of a set of 'items' and 'properties' that are used to describe features of those items.

A Wikidata item is defined by a unique identifier (e.g. Q131065), a label (i.e. *Chlamydia trachomatis*), a description (i.e. 'species of prokaryote') and a set of 'claims' about the item organized into 'statements' (Figure 2). A statement consists of a triple with an item as the subject, a Wikidata-defined property as the predicate (i.e. taxon rank, property P105), and another Wikidata item or Literal data value as

the object. Optionally, a set of references can be added as evidence and provenance for the claim made by the triple, and qualifiers can specify the context where the claim is valid (https://www.mediawiki.org/wiki/Wikibase/Data Model/Primer).

The 'ontology' of Wikidata is determined by the set of properties that may be used to create claims about the items within it. Entities can be created at any time, but properties can only be created by elected administrators following community discussion. The properties needed to support our current data model are listed in Table 1. It is worth noting that most of these properties are either generic (i.e. subclass of) or were defined by the Molecular Biology WikiProject (https://www.wikidata.org/wiki/Wikidata:WikiProject_Molecular_biology) before the outset of the present work on microbial genomes. Likewise, many requisite items, e.g. entities representing drugs and enzymes, were also present at the outset of the project and continue to be improved by members of the community.

Figure 2. Example Wikidata item. A Wikidata item defined by its label, description and statements that provide annotations and create relationships with other items in the database. (https://www.wikidata.org/wiki/Q131065).

Table 1 Wikidate properties

ID	Name	Value type
P685	NCBI taxonomy ID	String
P105	Taxon rank	Wikidata item
P171	Parent taxon	Wikidata item
P2249	RefSeq genome ID	String
P1542	Cause of	Wikidata item
P351	Entrez gene ID	String
P279	Subclass of	Wikidata item
P703	Found in taxon	Wikidata item
P644	Genomic start	String
P645	Genomic end	String
P702	Uniprot ID	String
P637	RefSeq protein ID	String
P702	Encoded by	Wikidata item
P688	Encodes	Wikidata item
P680	Molecular function	Wikidata item
P681	Cellular component	Wikidata item
P682	Biological process	Wikidata item
P361	Part of	Wikidata item
P128	Regulates	Wikidata item
P1056	Product	Wikidata item
P2175	Medical condition treated	Wikidata item
P2176	Drug used for treatment	Wikidata item

Already, Wikidata is showing how an open system can evolve over time with subsequent efforts building directly on prior work.

Some of the general purpose properties such as 'product', 'part of', 'cause of' and 'regulates' are currently used to establish the connections in Figure 1, but are likely to be replaced or extended with more biology-specific relations (such as 'precursor' and 'substrate of') over time. The other aspects of the current model that are more specific to representing microbial data are depicted in Figure 3.

One key requirement for modeling microbial data is the capacity to represent multispecies, multi-strain datasets. Microbiome and genomic research require the ability to do both intra- and interspecies comparative analysis. To support this work, our model follows a hierarchical taxonomy ranking scheme with the microbial species assigned to a Wikidata item (i.e. *C. trachomatis* #Q131065) defined by the core properties 'NCBI Taxonomy ID' (P685) ('813'), 'Taxon Rank' (P105)('species') and 'Parent Taxon' (P171)('*Chlamydia*').

Since genome annotations are based on the genome assembly of the specific strain sequenced and that genome assembly has its own unique identifier (i.e. NCBI RefSeq Genome Accession number), strain level distinction is
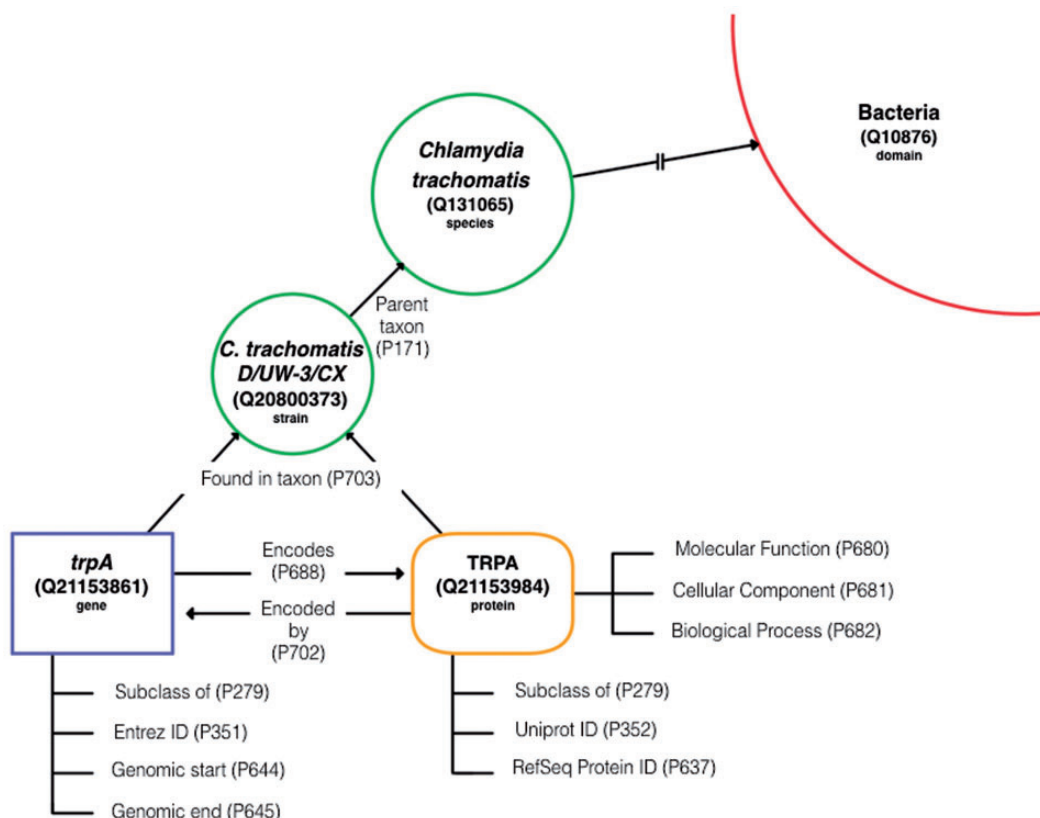
Figure 3. Data model template. The basic framework of the microbial genetic data model in Wikidata showing items and the statements that connect them. Item types are demarcated by label and color (i.e. gene item = blue and protein item = orange).

critical in bacteria. Individual strain items (e.g. *C. trachomatis* D/UW-3/CX #Q20800373) include the same core properties as a species item, the 'RefSeq Genome ID' (P2249), and are linked to the species item via the 'Parent taxon' (P171) property.

On the molecular level, the gene and protein must be kept as distinct entities, while maintaining their connections for queries down the line. A microbial gene item contains the similar core properties of a human gene item, including 'Entrez Gene ID' (P351) and 'Subclass of' (P279), but, 'Found in taxon' (P703) was added to distinguish which strain/genome assembly this particular gene came from. The gene links to its product item via the 'Encodes' (P688) property and reciprocally, the protein item will link to the gene that encoded it by the 'Encoded by' (P702) property. Core properties for microbial protein include 'RefSeq Protein ID' (P637), 'UniProt ID' (P352), 'Found in taxon' (P703) and 'Subclass of' (P279). Functional annotations are downloaded from the UniProt protein record and included here as subclasses of the gene ontology terms, 'molecular function' (P680), 'cellular component' (P681) and 'biological process' (P682)'.

Populating and querying microbial genes in Wikidata

Given the data model depicted in Figure 3 and encapsulated in the properties listed in Table 1, we have seeded Wikidata with representative content for 21 species of bacteria totaling 43 694 gene and 37 966 protein items, from various public databases. This work was carried out with a 'bot', a program for making automated edits in Wikidata, with source code available at www.bitbucket.org/sulab/wikidatabots/src. In addition, we manually established all of the Wikidata items and relationships needed to realize the operon data structure in Figure 1. This information can be accessed through the various APIs offered by Wikidata (https://www.wikidata.org/w/api.php, https://query.wikidata.org/). As an example, a user can easily retrieve all genes, proteins, and gene ontology annotations for the two strains of Chlamydia that are currently loaded using a Wikidata SPARQL query (Figure 4).

Note that the query actually requests this information for all bacteria through the '?taxa wdt:P171* (parent taxons) wd:Q10876 (Bacteria)' aspect of the query. As more bacterial genomes are loaded by us or other groups, the same query will return more and more data.

As another example, the following SPARQL query returns all operons, their regulators and their products (Figure 5).

Revisiting the example question illustrated in Figure 1 regarding organisms that are likely to be related to the persistence of chlamydial infections, we can ask what microbes are located in the female urogential tract and capable of generating indole (Figure 6). This query returns *E. coli* IAI39, a bacterial strain commonly found in patients with a perturbed vaginal microbiome (7). This query implicates an additional microbe suspect in chlamydial evasion of the host immune response. With disease–drug mappings in Wikidata, imported from the **National Drug File - Reference Terminology (NDF-RT)** database (17), queries that identify antibiotic compounds with action against *E. coli*, *C. trachomatis* or both are possible (Figure 7). These queries could inform an alternative treatment strategy that also targets the indole-producing bacteria in the vaginal microbiome, as was suggested by the original authors of the hypothesis (6).

Discussion

Wikidata is certainly not a replacement for core data curation centers such as NCBI and UniProt. But it could form the basis for a complementary, stable and cost-effective approach for capturing content that is either left trapped in the literature or represented only in small databases subject to the perils of funding cuts and general link rot. Though the microbial queries listed above currently return only a small fraction of the relevant content that exists in the world, the power of the Wikidata approach is that our seedling database can be extended by anyone with the will to do so.

Wikidata is now edited by > 15 000 active users and currently has over 15 million content pages (https://www.wikidata.org/wiki/Special:Statistics). Because of its open structure, its change tracking features and its evidence-capturing data model, it encourages community participation at all levels. While the community consensus building process can be slow and at times frustrating, it drives the stability and quality of Wikidata content.

Topic-specific databases can lose funding and disappear (18). Even government-backed institutions like NCBI and European Bioinformatics Institute (EBI) (http://www.ebi.ac.uk/) are vulnerable to funding cuts depending on the political climate. The unique connection between Wikidata and all the Wikipedias already make it one of the most well-known and easily discoverable knowledge bases in the world (19, 20). Data deposited here is far less likely to be lost, especially when care is taken to weave it into what already exists. Every item loaded to Wikidata (MediaWiki Foundation's third most active project) becomes a fixed point in a stable, self-sustaining knowledge representation platform that anyone can add to, and anyone can help the network grow through sharing the benefits of their own expertise.

```
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>
SELECT DISTINCT ?taxa_name ?gene_id ?uniprot_id ?prot_name ?go_name
WHERE {
?gene wdt:P351 ?gene_id ;
wdt:P688 ?protein ;
  wdt:P703 ?taxa .
  ?protein rdfs:label ?prot_name ;
    wdt:P352 ?uniprot_id ;
    ?function_type ?go_term .
  ?go_term rdfs:label ?go_name .
  ?taxa wdt:P171* wd:Q10876 ;
    rdfs:label ?taxa_name .
FILTER (LANG(?go_name) = "en") .
FILTER (LANG(?taxa_name) = "en") .
FILTER (LANG(?prot_name) = "en") .
}
```

taxa_name	gene_id	uniprot_id	prot_name	go_name
Chlamydia trachomatis 434/BU	5858781	A0A0H3MCU1	hypothetical protein CTL0276	protein
Chlamydia trachomatis 434/BU	5857942	B0B8K9	Na(+)-translocating NADH-quinone reductase subunit A NQRA	protein
...	...	...	...	...
Chlamydia trachomatis 434/BU	5858751	A0A0H3MDG2	4-hydroxy-tetrahydrodipicolinate synthase DAPA	protein
Chlamydia trachomatis 434/BU	5858405	A0A0H3MLC8	replicative DNA helicase DNAB	protein

Figure 4. SPARQL query for all microbial genes, proteins and associated Gene Ontology annotations in Wikidata. Properties used: P351 = entrez_gene_id, P688 = encodes, P703 = found in taxon, P352 = uniprot_id, P171 = parent taxon. Note that the * operator on P171* results in a recursive search for organisms that descend from wd:Q10876 (Bacteria). This query may be executed at https://query.wikidata.org/.

The open access, community-driven nature of Wikidata contributes to its perpetuity, but the major enduring factor is its universal utility. Wikidata is a place for knowledge of any conceivable topic from surfing (Q159992) to bacteria (Q10876). This variety of topics generates community support that a domain-specific, funding-dependent database cannot compete with. In addition to support, Wikidata creates the ability to link a surfboard (Q457689) to surfing (Q159992), the 'sport' (P641) it is used in, or *C. trachomatis* D/UW-3/CX (Q20800373) to pelvic inflammatory disease (Q558070), a disease it is the 'cause of' (P1542) in humans (Q5). Moreover, in principle, it allows microbial genetics data to be linked to data from related fields, including pharmacology and epidemiology.

These relationship examples highlight another powerful virtue of Wikidata compared to other data storage platforms; adding data to Wikidata requires the use of meaningful properties for relating entities. It is insufficient to simply state 'rdf:seeAlso' as the link between two related

```
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>
SELECT ?taxa_name ?regulator_name ?operon_name ?go_name ?product_name
WHERE {
?operon wdt:P279 wd:Q139677 ;
rdfs:label ?operon_name ;
    wdt:P527 ?gene ;
    wdt:P1056 ?protein .
?regulator wdt:P128 ?operon  ;
    rdfs:label ?regulator_name .
?protein ?function_type ?go_term ;
    wdt:P1056 ?product .
?go_term wdt:P686 ?go_id ;
    rdfs:label ?go_name .
?product rdfs:label ?product_name .
?gene wdt:P703 ?taxa .
?taxa rdfs:label ?taxa_name .
FILTER (LANG(?taxa_name) = "en") .
    FILTER (LANG(?regulator_name) = "en") .
    FILTER (LANG(?go_name) = "en")
    FILTER (LANG(?product_name) = "en") .
}
```

taxa_name	regulator_name	operon_name	go_name	product_name
Chlamydia trachomatis 434/BU	Trp operon repressor trpR	trpRBA operon	tryptophan synthase activity	Water
Chlamydia trachomatis D/UW-3/CX	Trp operon repressor trpR	trpRBA operon	tryptophan synthase activity	Water
...	...	...	...	...
Chlamydia trachomatis D/UW-3/CX	Trp operon repressor trpR	trpRBA operon	tryptophan synthase activity	L-Tryptophan
Chlamydia trachomatis D/UW-3/CX	Trp operon repressor trpR	trpRBA operon	tryptophan synthase activity	L-Tryptophan

Figure 5. SPARQL query for all operons, their regulators, the taxon that expresses them and their functional products in Wikidata. Q139677 is the Wikdata item for the class 'operon'. Properties used: P279 = subclass of, P527 = has part, P1056 = product, P128 = regulates, P688 = encodes, P703 = found in taxon. This query may be executed at https://query.wikidata.org/.

entities (as many major databases do in their Resource Description Framework (RDF) (https://www.w3.org/RDF/) representations). A relationship between items cannot be added without an appropriate property in place, requiring the data model to be defined prior to importing the data. This process of creating properties through community discussion and consensus drives the development of their ontology up-front, rather than forcing the burden of integrating ambiguous content downstream to consumers.

Limitations

Although Wikidata provides an excellent framework for housing some forms of data, it has some limitations. Not all content is appropriate for Wikidata. It is a database of referenced claims about the world and should not, for example, be a repository for sequences or raw expression data. There is no built-in reasoning in Wikidata. Unlike other Semantic Web implementations, e.g. those built using the OWL ontology language (21), the current

Wikidata system cannot be used to infer new facts nor to automatically check the logical consistency of statements. Editors cannot be constrained from making claims that may break data models spread across multiple items. As an openly editable resource, it is possible for data to be disrupted by edits from both well-intentioned editors and, at least theoretically, by malicious users (though true vandalism has thus far not happened at detectable levels). In practice, system-level limitations on data integrity checking are dealt with through the operation of user-generated 'bots' that patrol wikidata seeking and correcting problems as they arise. In terms of inferring new facts based on what is represented in wikidata, it is possible to load all of its contents into external implementations that support this kind of reasoning via frequently updated exports of the entire database (https://www.wikidata.org/wiki/Wikidata:Database_download).

Apart from limits on what content can be stored and what operations can be applied to that content, wikidata is also constrained in terms of user interface. When users browse items online at https://wikidata.org, they are seeing one interface for viewing and editing items ranging from protein domains to Dutch history museums. While it is possible to use this interface to find useful information, e.g. all the connections that form the network of knowledge depicted in Figure 1 can be clicked through, it is certainly not the most ideal human interface for this task. For developers, the SPARQL query interface provides the most

```
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>
SELECT ?organism_name WHERE {
  ?organism_item wdt:P276 wd:Q5880 ;
    rdfs:label ?organism_name .
  ?gene wdt:P703 ?organism_item ;
    wdt:P1056 wd:Q319541 .
  FILTER (LANG(?organism_name) = "en") .
}
```

organism_name
Escherichia coli IAI39

Figure 6. SPARQL query for all organisms that are located (P276) in the female urogential tract (wd:Q5880) and that have a gene with product (P1056) indole (wd:Q319541). This query may be executed at https://query.wikidata.org/.
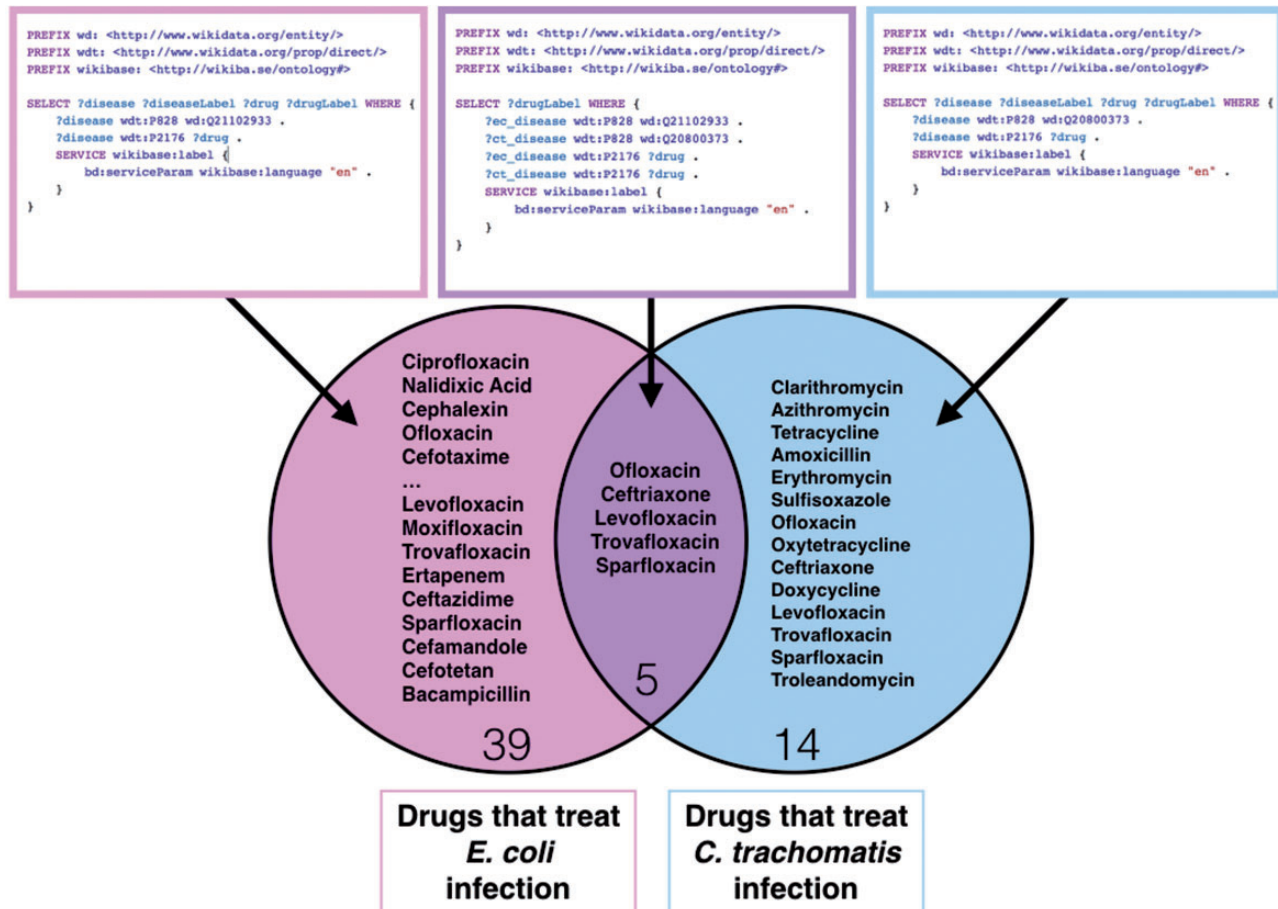

Figure 7. SPARQL query for all drugs used for treatment (P2176) for commensal *E. coli* infectious disease (Q18975220), *Chlamydia* infection (Q153356) and both *E. coli* and *Chlamydia* infections with a venn diagram display of the results. These queries may be executed at https://query.wikidata.org/.

effective way to access data. For end-users, domain-specific applications need to be written that expose the content within Wikidata. These applications will support both effective information retrieval and visualization, but can also support direct curation of the content by domain experts.

Next steps

Even in consideration of these limits, Wikidata is a tremendous potential platform for managing the process of collaboratively understanding microbial genomics. In support of this objective, our immediate next steps are to load the remaining 99 microbial reference genomes from NCBI (http://www.ncbi.nlm.nih.gov/genome/browse/reference/), encompassing bacteria that are the most studied and relevant to human health. This will load an additional ∼900 000 entities between genes and gene products. These items will form a foundation upon which we invite the microbial research community to collaboratively synthesize their knowledge as it evolves into the future. In addition to loading this data, we are collaborating on the development of an extension to the Web Apollo genome browser that will allow genome annotations from Wikidata to be viewed and edited directly within this popular tool (22). We see this as the first of many bioinformatics applications that can be enhanced by Wikidata content and can enable expert users to contribute to its collaborative curation.

Conclusion

We invite and encourage the rest of the scientific community to join our cause in creating this universal microbial genomics resource. We have shown that the aggregation of a subset of data enables powerful queries that demonstrate the potential of connecting data from fragmented sources into a centralized, well-defined structure. In addition to data that can be collected from structured data sources and aggregated in Wikidata by bot, a great deal of important information resides only in primary literature. Accessing and integrating this content requires human editors. It is thus imperative that we engage the microbial research community to help build Wikidata to its potential; a centralized, Semantic Web compatible mechanism for making sense of mountains of microbial genetic data.

Acknowledgements

Funding

References

1. Gilbert,J.A., Jansson,J.K. and Knight,R. (2014) The Earth Microbiome project: successes and aspirations. *BMC Biol.*, 12, 69.
2. Tatusova,T., Ciufo,S., Fedorov,B. *et al.* (2015) RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.*, 43, 3872.
3. UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res*, 43, D204–D212.
4. Okuda,S. and Yoshizawa,A.C. (2011) ODB: a database for operon organizations, 2011 update. *Nucleic Acids Res.*, 39, D552–D555.
5. Akers,J.C. and Tan,M. (2006) Molecular mechanism of tryptophan-dependent transcriptional regulation in *Chlamydia trachomatis*. *J. Bacteriol.*, 188, 4236–4243.
6. Caldwell,H.D., Wood,H., Crane,D. *et al.* (2003) Polymorphisms in Chlamydia trachomatis tryptophan synthase genes differentiate between genital and ocular isolates. *J. Clin. Invest.*, 111, 1757–1769.
7. Gibbs,R.S. (1987) Microbiology of the female genital tract. *Am. J. Obstet. Gynecol.*, 156, 491–495.
8. Aiyar,A., Quayle,A.J., Buckner,L.R. *et al.* (2014) Influence of the tryptophan-indole-IFNγ axis on human genital *Chlamydia trachomatis* infection: role of vaginal co-infections. *Front. Cell. Infect. Microbiol.*, 4, 72.
9. Law,V., Knox,C., Djoumbou,Y. *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, 42, D1091–D1097.
10. Chang,A., Schomburg,I., Placzek,S. *et al.* (2015) BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic Acids Res.*, 43, D439–D446.
11. Macklaim,J.M., Clemente,J.C., Knight,R. *et al.* (2015) Changes in vaginal microbiota following antimicrobial and probiotic therapy. *Microb. Ecol. Health Dis.*, 26, 27799.
12. Vejborg,R.M., Hancock,V., Schembri,M.A. *et al.* (2011) Comparative genomics of Escherichia coli strains causing urinary tract infections. *Appl. Environ. Microbiol.*, 77, 3268–3278.
13. Howe,D., Costanzo,M., Fey,P. *et al.* (2008) Big data: the future of biocuration. *Nature*, 455, 47–50.
14. Huss,J.W.,3rd, Orozco,C., Goodale,J. *et al.* (2008) A gene wiki for community annotation of gene function. *PLoS Biol.*, 6, e175.
15. Good,B.M., Clarke,E.L., Loguercio,S. *et al.* (2012) Linking genes to diseases with a SNPedia-Gene Wiki mashup. *J. Biomed. Semantics*, 3 (Suppl 1), S6.
16. Robinson, I., Webber, J., and Eifrem, E. (2013) *Graph Databases*. O'Reilly Media, Inc. , Sebastopol, California, USA.
17. Carter,J.S., Brown,S.H., Bauer,B.A. *et al.* (2006) Categorical information in pharmaceutical terminologies. *AMIA Annu. Symp.*

From proceedings of an annual symposium. http://www.ncbi.nlm.nih.gov/pubmed/17238314

18. Wren,J.D. (2008) URL decay in MEDLINE–a 4-year follow-up study. *Bioinformatics*, 24, 1381–1385.

19. Vrandečić,D. and Krötzsch,M. (2014) Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57, 78–85.

20. Elvira, M., Andra, W., Sebastian, B.L.M. *et al*. (2015) Wikidata: a platform for data integration and dissemination for the life sciences and beyond. Proceedings of the 2015 Swat4LS International Conference in Cambridge England. http://ceur-ws.org/Vol-1546/

21. McGuinness,D.L. and Van Harmelen,F. (2004) OWL web ontology language overview. *W3C Recommendation*, 10

22. Lee,E., Helt,G.A., Reese,J.T. *et al*. (2013) Web Apollo: a web-based genomic annotation editing platform. *Genome Biol.*, 14, R93.

23. Zubáčová,Z., Krylov,V. and Tachezy,J. (2011) Fluorescence in situ hybridization (FISH) mapping of single copy genes on Trichomonas vaginalis chromosomes. *Mol. Biochem. Parasitol.*, 176, 135–137.