PPSDB : A Linked Open Data knowledge base for protist-prokaryote symbiotic interactions

Brandon K. B. Seah (http://orcid.org/0000-0002-1878-4363)
Thünen Institute for Biodiversity, Bundesallee 65, 38116 Braunschweig, Germany.
Email: brandon.seah@thuenen.de kb.seah@gmail.com

Abstract

As the ecological and evolutionary importance of symbiotic interactions between protists (microbial eukaryotes) and prokaryotes (bacteria and archaea) is better appreciated, keeping an overview of their diversity and the literature becomes a growing and ongoing challenge. Here we present the Protist-Prokaryote Symbiosis Database (PPSDB), comprising 789 manually curated interaction statements sourced from 410 publications, where biological taxonomy, anatomical localization, and analytical methods applied have been annotated and mapped to external databases and ontologies, such as Wikidata, NCBI Taxonomy and Gene Ontology. We describe how our data model deals practically with challenges such as incomplete information and inconsistent taxon concepts, which will be applicable to similar projects. Both the model and underlying Wikibase software platform are highly extensible, so new items and properties can easily be added. Unlike a static table or list of citations, PPSDB is a structured knowledge base that enables programmatic access and powerful, integrated semantic queries. The database is available at https://ppsdb.wikibase.cloud/.

Keywords: linked open data, knowledge graph, FAIR, SPARQL, semantic web, Wikibase.cloud

Introduction

The study of protists (microbial eukaryotes) has revealed a fascinating diversity of interactions with prokaryotes, including symbionts that defend their hosts, help them move, and orient them in the environment (Petroni et al. 2000; Hongoh et al. 2007; Monteil et al. 2019). New symbioses are regularly discovered, e.g. by mining protist genome data for prokaryote sequences (Davison, Hurst, and Siozios 2023), while previously described ones can be profitably revisited with modern methods. To contextualize new discoveries, and spot larger trends and knowledge gaps, an accurate overview of the current body of knowledge is indispensable.

The existing entry points to knowledge on the diversity of protist-prokaryote interactions are research articles and reviews, e.g. (Ball 1969; Bjorbækmo et al. 2020; Husnik et al. 2021; Kostygov et al. 2021; Shi et al. 2021; Fokin and Serra 2022). These are static, usually not interlinked with primary data, and formatted for human readers rather than programmatic queries, even if some information is presented in tabular form. It is therefore difficult to get a reliable answer to a question such as "what protists have Alphaproteobacteria symbionts localized in the host nucleus" without a deep dive into the literature for oneself.

The study of biotic interactions also suffers from poor discoverability of relevant information (Poelen, Simons, and Mungall 2014; Mihara et al. 2016). Ideally, databases should be more than just a list of taxa or citations, but also capture other facets such as phylogenetic affiliation, interaction type, and environmental context. To achieve this, they should be built on a semantic data model that is extensible enough to accommodate new information, concepts, and terminology as they arise.

Specific challenges are posed by the ever-evolving biological taxonomy and methods used to identify and describe organisms. Better taxon sampling and methods constantly drive updates to the higher taxonomy and nomenclature of both eukaryotes (Adl et al. 2019) and prokaryotes (Parks et al. 2020). For taxa originally identified or described on the basis of morphology or phenotype, their placement within modern taxonomies may be unclear, although some symbioses have been revisited with sequencing to clarify their phylogenetic identity (Boscaro et al. 2013; Schrallhammer, Castelli, and Petroni 2018). In contrast, many recent studies assign names solely from sequence data. Many organisms, particularly environmental microbes, never receive a formal scientific name and remain known under an informal or provisional name even if otherwise well characterized. Therefore, in addition to

names and taxonomy, the analytical methods and evidence base behind each described symbiotic interaction should also be documented.

For the uses envisioned above, we argue that the information is best managed in the form of a knowledge graph. Knowledge graphs are data structures that represent concepts/entities and the relationships between them abstractly as the nodes and edges of a directed graph (network) (Chaudhri et al. 2022). A common way to specify a graph's structure is as a collection of linkages, each comprising two nodes and the edge that connects them. The meanings assigned to nodes and edges depend on the domain-specific application of the knowledge graph: for symbioses, nodes can represent biological taxa, and edges their interactions. Complex multi-way or nested relationships as well as incomplete information can thus be represented more naturally and efficiently than in a tabular format or relational database.

Database items can be assigned standardized identifiers (uniform resource identifier, URI), and be further linked ("mapped") to equivalent entities in other databases, e.g. for taxa or publications. URIs can be a web address that returns useful information about the entity. Data published following such principles are known as Linked Open Data (LOD) (Bauer and Kaltenböck 2012). Taxon names can be seen as an analogy to URIs within biology, as they also aim to be (ideally) unique and language-agnostic identifiers for real-world entities. Taxon names remain the primary vehicles through which biologists convey and retrieve information about organisms, albeit intended for human use and recall (Patterson et al. 2010). Machine-readable URIs fulfill a similar role but can be processed programmatically. Linking equivalent concepts or entities between datasets with URIs lets us build on other databases and avoid duplicated effort; for example, we do not need to curate a full biological taxonomy in our own database if we map taxa to an existing, programmatically accessible taxonomic database. Other concepts/entities, such as anatomical and environmental terms, can be mapped to ontologies, such as the Gene Ontology (Gene Ontology Consortium et al. 2023) and Environment Ontology (Buttigieg et al. 2016). This not only ensures that terminology is used consistently with the wider community, but also allows sophisticated queries that take advantage of the semantic relationships encoded in those ontologies (Pacheco et al. 2022).

Here, we describe a knowledge base for protist-prokaryote symbiotic interactions, and showcase how Linked Open Data principles enable powerful, integrated searches across multiple resources. The design objectives were to: (i) Represent curated information from the scientific literature, with citations for each statement. (ii) Focus on named symbiotic

interaction partners from low-diversity systems, rather than microbiome studies dealing with higher level taxa or OTUs. (iii) Link records to sequence databases. (iv) Enable multiple entry points for queries, including biological taxonomy, anatomical localization of symbionts, and analytical methods used to identify organisms. (iv) Map concepts and entities in the database to external taxonomies, ontologies, and identifiers, to ensure that they are described consistently and interoperable with other resources and knowledge representations.

Methods

Software platform and tools

The database was built on a Wikibase instance hosted by Wikibase.cloud, a service provided by Wikimedia Deutschland. The database was edited through the web interface, through batch edits using the QuickStatements tool (https://www.wikidata.org/wiki/Help:QuickStatements), and programmatically with Python scripts (https://github.com/kbseah/ppsdb-utils/) using the WikibaseIntegrator library v0.12.5 (https://github.com/LeMyst/WikibaseIntegrator). The data dump to XML was performed with mediawiki-dump-generator (https://github.com/mediawiki-client-tools/mediawiki-dump-generator). Periodic exports for indexing by Global Biotic Interactions (GloBI) are hosted on GitHub (https://github.com/kbseah/ppsdb-globi-export).

Data model and terminology

In the Wikibase platform, nodes are called "items", and the edges connecting them are assigned specific meanings, or "properties". Each connection (two items linked by a property) makes up a "statement", with one item as the subject and the other as the object (Figure 1). Statements themselves can be treated like items and be the subject of further "qualifier" statements that provide additional information. Statements can also be annotated with references, which are a special type of qualifier statement.

Items may belong to one of two types, "classes" and "instances", following the usage in Wikidata (https://www.wikidata.org/wiki/Help:Basic_membership_properties). A class is a set of items that have common properties; the members of a class are known as instances. Classes may be further subdivided into subclasses. For example, "*Pelomyxa palustris*" is an instance of the class "formally named taxon", which is a subclass of "taxon". Other classes in PPSDB represent references, organismal body parts, analytical methods, environmental terms, and interaction types; all items are ultimately descended from the root class "entity".

We modeled each biotic interaction as a statement linking two taxon items with an "interacts with" property (Figure 1, Figure 2). Each statement was further qualified with (i) where the symbiont is localized in the host organism/cell, (ii) the analytical methods used to identify (taxonomically and phylogenetically) host and symbiont, and (iii) the nature of the biotic interaction (e.g. transfer of fixed organic carbon, pathogenic), if known. Further statements on taxon items mapped them to external taxonomy databases and representative sequence records, and described the environmental context from which organisms were isolated or sampled (Figure 1, Figure 2).

Data collation and mappings to external identifiers

Reported symbiotic interactions between protists and prokaryotes were gathered from the published literature through ad hoc keyword searches and relevant review articles. These included studies that specifically focused on symbiosis, as well as morphological or taxonomic studies that incidentally described associated microbes.

Relevant information was extracted from original research publications where possible, and mapped to external identifiers if a suitable exact match existed (Table 1). Taxon items were created to represent the interacting organisms; these are understood to be at species rank or below, even if they are only identified to a higher ranked taxon, similar to the concept of a "submittable taxon" in ENA (https://ena-docs.readthedocs.io/en/latest/faq/taxonomy.html). This was to avoid potential conflation of multiple taxa under the same identifier. If sequence data were available, the taxon was linked to a representative sequence named in the publication describing it, to ensure that the name is associated with empirical data should disambiguation be necessary in the future. Equivalent identifiers in the NCBI Taxonomy (Schoch et al. 2020) were linked if available. Formally described or *Candidatus* (for prokaryotes) taxa were linked to Wikidata and the List of Prokaryotic Names with Standing in Nomenclature (LPSN) (Parte et al. 2020) on the basis of taxon name.

The localization of symbionts in the host was mapped to cellular anatomy terms in the Gene Ontology (GO) (Ashburner et al. 2000; Gene Ontology Consortium et al. 2023) or metazoan anatomical terms in Uberon (for protists that are also symbionts of animals) (Mungall et al. 2012). The relationships between anatomical terms are represented in ontologies, which can be exploited when performing queries.

Cited publications were linked to digital object identifiers (DOIs) and to Wikidata, where bibliographic data are maintained by a community project, Wikicite. Citations missing from Wikidata are easily imported with the Scholia tool (Nielsen, Mietchen, and Willighagen 2017). Formatted citations were obtained from CrossRef from their DOIs, otherwise added manually.

Three properties were used to describe the environmental context at different scales– broad scale environmental context, local environmental context, and environmental material– using Environment Ontology (EnvO) terms (Buttigieg et al. 2013, 2016), following the MIxS guidelines (https://github.com/EnvironmentOntology/envo/wiki/Using-ENVO-with-MIxS) (Yilmaz et al. 2011). If equivalent terms existed, analytical techniques were mapped to the Ontology for Biomedical Investigations (OBI) (Bandrowski et al. 2016), and interaction types were mapped to the OBO Relations Ontology (https://github.com/oborel/obo-relations).

Challenges for data mapping and modeling

Most challenges related to taxonomy, because informal or provisional names are often used for microorganisms, and disparate methods and levels of detail have been used to characterize them. Ideally, each organism would be described in a scientific publication under a formal taxon name, accompanied by published sequence accessions and an equivalent NCBI taxon item. A new Wikidata item for the taxon was created if it did not already exist. The PPSDB item was then linked to the corresponding Wikidata and NCBI Taxonomy items. Here we describe how we dealt with other cases that were not so neatly organized.

Informal or provisional taxon names with NCBI Taxonomy equivalent

Many studies described organisms without assigning a formal taxon name, but their phylogenetic affiliation is nonetheless known, sequence data are available, and the corresponding informally named taxon item in the NCBI Taxonomy appears to be equivalent to the taxon concept in the study. If so, the item was mapped to that NCBI taxon ID, and labeled with the informal name used in the cited publication, as well as known aliases from other publications and databases.

Taxon concept not in NCBI Taxonomy, but sequence data available

Taxon items in the NCBI Taxonomy and taxonomic annotations of sequence records may not be up to date, or may differ from the published literature. For example, the ciliate species *Eufolliculina methanicola* (https://ppsdb.wikibase.cloud/entity/Q52) was formally described in a scientific publication (Pasulka et al. 2017), but sequences from that study were published

in Genbank under a placeholder taxon "Folluculinidae sp." (NCBI:txid1934002), which is used in the NCBI Taxonomy for records which were only identified to the family level and so may represent a mixture of different species. The PPSDB item was therefore not mapped to the NCBI Taxonomy, because it can lead to incorrect results if the identifier is used to retrieve sequence data. For this example, a formal taxon name was published, so a Wikidata item was created for it and mapped to PPSDB. To allow us to track the identity, should the NCBI Taxonomy be updated in the future, a representative SSU rRNA sequence record for this taxon that was cited in the original publication (KX012915) was linked to this item with the property "representative SSU rRNA sequence record".

Taxon name/concept with no sequence data available

An organism may have been identified by morphology alone, without using sequencing methods, or sequence data produced in a study cannot be found. For example, a species of *Arcobacter* (https://ppsdb.wikibase.cloud/entity/Q410) was identified as a symbiont of *Bihospites bacati* (https://ppsdb.wikibase.cloud/entity/Q409), but although sequencing of a marker gene was reported, the sequence was not published. Alternatively, the organism may have been identified to a higher taxonomic group by its morphology or with methods such as group-specific molecular probes. For example, the ciliate *Frontonia leucas* (https://ppsdb.wikibase.cloud/entity/Q1782) is associated with an unclassified Alphaproteobacteria (https://ppsdb.wikibase.cloud/entity/Q158) that was identified with group-specific molecular probes. As no direct sequence data were available in these cases, there were no sequence records to anchor the taxon concept empirically (see above). PPSDB items for such taxa were labeled with descriptive names based on what was reported in the publication, e.g. "unclassified Alphaproteobacteria", but were not mapped to external identifiers. Similarly named *incertae sedis* items may exist in the NCBI Taxonomy (e.g. "unclassified Alphaproteobacteria", NCBI:txid33807), but these were deliberately not mapped from PPSDB because they will pull in incorrect results if used for programmatic queries.

Consistent mapping of items to higher taxonomy

Not all organisms described in the literature could be mapped to external taxonomies, nor were they always identified to the species level. Nonetheless, we linked all taxon items to the next-highest-ranking, formally named parent taxon that was represented in both Wikidata and NCBI Taxonomy. This enabled consistent searches by taxonomy, even if the species

themselves were not mapped to an external taxonomy. The parent taxon items are instances of a class "higher taxon" that is not used in interaction statements.

Experimentally induced interactions

A number of symbiotic microbes were first identified in one host species but maintained in the laboratory in a different host because the original host was not suitable for experiments. For example, *Acanthamoeba castellanii* has been used as a lab host for various intracellular bacteria isolated from other amoebae (Schulz et al. 2016). These experimentally induced interactions were represented with a different property, "interacts experimentally with", to distinguish them from naturally occurring interactions.

Results and Discussion

PPSDB is hosted by Wikibase.cloud and is browsable through the web interface at https://ppsdb.wikibase.cloud/. A SPARQL endpoint is available for programmatic queries, with examples to help users get started: https://ppsdb.wikibase.cloud/query/. The structured data (in the Item: and Property: namespaces) are released under a CC0 1.0 public domain dedication (https://creativecommons.org/publicdomain/zero/1.0/).

Database statistics

The database currently (16 Aug 2024) documents 789 biotic interactions between 498 host taxa and 514 symbiont taxa, with 410 references cited. The number of citations is incidentally similar to the 328 works cited by Gordon H. Ball in his 1969 review, "Organisms living on and in Protozoa" (Ball 1969). However, given our focus on phylogenetic identity of the symbiotic partners and linking them to sequence data, there is a bias towards more recent publications in PPSDB and the overlap between the two sets of citations is minimal.

The most commonly represented host protist phyla are Ciliophora (136 taxon items), Metamonada (78), and Amoebozoa (55), while the most commonly represented symbiont prokaryotic phyla are *Pseudomonadota* (207 items), *Bacteroidota* (53), and *Methanobacteriota* (36). This undoubtedly reflects the activity of researchers rather than the abundance or ecological significance of these organisms. Some non-protist, non-prokaryote taxa are represented, e.g. termite hosts of metamonad flagellates that themselves have bacterial symbionts, as the host species helps to identify the flagellate. Multipartite interactions, or highly nested ones, are easily modeled, e.g. the bacterial epibionts of spirochaete ectosymbionts of flagellates from termite guts (Utami et al. 2019).

Usage examples

The SPARQL query engine bundled with Wikibase and the mappings to other databases, particularly Wikidata, allow users to query PPSDB in ways that are not easily implemented with other databases of similar scope. Further example SPARQL queries are listed at https://ppsdb.wikibase.cloud/wiki/Project:SPARQL/examples.

1. Symbionts localized in the host nucleus and their class (https://tinyurl.com/25fxfny9)

This search showcases how a query can use semantic content in the database and integrate an external database in the search.

The localization of intranuclear symbionts may be reported simply as "nucleus", or more specifically as "nuclear envelope lumen", or "macronucleus" (in ciliates, which have two developmentally distinct types of nuclei). The relationships between these terms are modeled in the database, e.g. nuclear envelope lumen is a part of, and macronucleus is a subclass of "nucleus". Users can be more or less specific as required.

We have chosen not to maintain a full biological taxonomy within PPSDB, but instead to map taxa to Wikidata and the NCBI Taxonomy (see "Challenges for data mapping and modeling" above). The search is therefore executed as a federated SPARQL query across both PPSDB and Wikidata. Most described intranuclear symbionts are *Alphaproteobacteria* from ciliates, but there are diverse hosts where the symbiont's phylogenetic position is unknown.

2. Interactions connected to *Ca.* Megaira polyxenophila (https://tinyurl.com/2948dund)

Ca. Megaira polyxenophila is known to associate with diverse host eukaryotes (Schrallhammer et al. 2013), and many of its known hosts in turn have more than one prokaryotic symbiont. These complex linkages can be visualized as a graph, which shows that some of these host species are also linked by other symbionts that have multiple host species: *Polynucleobacter necessarius* and *Caedimonas varicaedens* (Figure 3).

3. Symbionts identified by fluorescence in situ hybridization but not sequencing (https://tinyurl.com/282s455g)

A number of symbionts have been described in publications that employed group-specific molecular probes that could identify them to e.g. class level, but which did not sequence a phylogenetic marker gene, so a more precise classification was not possible. These may be interesting to revisit with modern sequencing methods.

4. Symbioses described in publications by a specific author (https://tinyurl.com/29gnxhw3)

We can query bibliographic metadata of the publications referenced in PPSDB such as authors and publication venues, via reference items mapped to Wikidata. Like the search by biological taxonomy, this is a federated SPARQL query. The search exploits the growing representation of publication and person data in Wikidata, which can be used for scientometric studies, such as investigating coauthorship networks.

Choice of software platform

We chose Wikibase as the platform for this database because it has both a web browser-based interactive interface and an API for programmatic access, and is available as a cloud service. Wikibase was originally developed as the backend for Wikidata, the largest open knowledge graph. As such, its design caters to the Wikidata model, but this potential limitation was outweighed by its ease of use, active user community, ongoing support and development, and integration of a SPARQL engine and other tools. Existing tools and libraries to work with Wikibase can be applied instead of reinventing the wheel. The ease of federated searches with Wikidata was also an advantage. Many current Wikibase users come from the cultural heritage field (Diefenbach, Wilde, and Alipio 2021; Huaman, Huaman, and Huaman 2023; Shimizu et al. 2023), and include institutions like the European Union and German National Library. PPSDB shows that an application in the natural sciences is straightforward.

Technical requirements and know-how remain a hurdle to the adoption of knowledge graphs. Wikibase.cloud is a good compromise for smaller projects and prototypes driven by subject-matter experts who may not have a deep background in semantic web technologies. No programming experience is required to get started, as data entry and editing can be performed through the web interface, with users learning additional tools (e.g. QuickStatements for tabular data entry, SPARQL for queries) as they go along. User management, project planning, and discussion pages can be maintained on the same wiki as the database itself, making it self-contained. Other projects with similar aims have built bespoke software, e.g. AQUASYMBIO http://www.aquasymbio.fr/, Viral Host Range DB https://viralhostrangedb.pasteur.cloud/ (Lamy-Besnier et al. 2021), and Virus-Host DB https://www.genome.jp/virushostdb/ (Mihara et al. 2016). Such software is harder to maintain in the long term, and requires more effort to integrate with other linked data sets. Similar considerations have been cited by Wikibase users who have migrated from other platforms (Koho et al. 2023).

Virtuous cycles of data curation

During data curation, we sometimes discovered outdated records or errors while mapping items to NCBI Taxonomy and Wikidata. We edited Wikidata directly, while the NCBI Taxonomy team was contacted by email with corrections. Commonly encountered issues included NCBI Taxonomy records that still used a provisional name although a formal taxon name or Candidatus name has been published, and taxon names or publications that were not yet represented on Wikidata. Linked open data naturally fosters collaboration and a mutualistic relationship between the linked resources, such that the curation and data cleaning of one benefits the others too (Seah 2023; von Mering et al. 2023).

Data sharing and archiving

Even if data are linked and open, interested users may not be able to find them easily. The core interaction data in PPSDB were therefore exported as a table for indexing by the Global Biotic Interactions (GloBI) database, an aggregator for species interaction data that is searchable from its web site and through an R package (Poelen, Simons, and Mungall 2014). This increases the visibility of protists, which are underrepresented in ecological studies, and of the original publications, which are cited in full. To secure the long term availability of the database, periodic XML dumps are archived on Internet Archive. The export for GloBI is also archived separately on Zenodo.

Remaining challenges for data modeling/mapping

Some types of statements in the literature remain difficult to represent formally in our data model. Broad statements about higher taxa, e.g., "all *Kentrophoros* species are associated with bacteria from genus *Candidatus* Kentron", are represented in our data model by creating individual items and links for each known species within those taxa. Such a statement implies that as-yet unstudied or undescribed host species will also be found to interact with corresponding symbiont species. However, we cannot create items for unknown species, so we conservatively do not add such implicit statements.

The modeling of facets other than biological taxonomy is relatively basic and can be further developed. For symbiont localization, we currently do not distinguish between different types of topological relationships. For example, methanogenic endosymbionts are typically located in the host cytoplasm close to hydrogenosomes, but this detail is not captured by the single "subject body part" property. For interaction types, most terms have not been mapped to the OBO Relations Ontology (RO), because many terms that are

meaningful to microbial ecologists, e.g. “syntrophy” and “auxotrophy”, do not appear in RO. The outcome (e.g. mutualistic vs. parasitic) and function of many microbial interactions is also unclear or only inferred. Nonetheless, more elaborate modeling of these aspects may be overly complex for most users, so a simpler representation may be more useful.

Microbiome survey studies were excluded from the scope of PPSDB. However, some larger protists are associated with diverse prokaryotes; some may be stable partnerships while others are facultative. Sequencing surveys may reveal dozens of such interactions per host species—should they all be included in the database?

Finally there are the practical hurdles we experienced when extracting relevant information from publications. Taxon names and sequence accessions may be scattered in different parts of a manuscript and its supplements or even across multiple publications, different names may be used for the same organism at different times, methods may be incompletely reported, and in a few (thankfully rare) cases, which symbiont belongs to which host was not reported at all even though both are separately characterized. We suggest that authors summarize biotic interaction results in tabular format where any sequence accessions or identifiers are also directly listed.

Future directions

We have described how we model biotic interactions in a knowledge graph, and our solutions to challenges such as taxonomic inconsistency, uncertainty, and the proliferation of names and identifiers. The model can easily be adapted to represent other types and facets of biotic interactions by adding new classes and properties. For example, a statement representing an allelopathic interaction between two plant species could have a qualifier that links to the phytochemical responsible, the latter represented as an item of class “chemical” and mapped to databases like ChEBI. PPSDB itself can be extended to encompass other taxonomic groups; viruses are particularly relevant as some giant viruses of protists were initially thought to be bacterial symbionts (La Scola et al. 2003). As mentioned above, a limited number of non-protist, non-prokaryote taxa are already represented in the database.

The current bottleneck in data curation is in discovering and parsing the relevant scientific literature. Other projects have data-mined molecular sequence metadata, e.g. the “host” tag in Genbank records, to produce large-scale species interaction datasets (Wardeh et al. 2015; Albrycht et al. 2022). Such pipelines are most suitable for taxa, such as viruses, that are routinely described with sequencing, where relevant metadata are generally accessible in

standardized form. We are however also interested in the historical literature, and in symbioses described with other methods, such as microscopy, where data deposition and metadata reporting are not yet as standardized as for sequencing. For the foreseeable future, we still need humans to verify that the data are reliable, but this could be supplemented by provisional interaction claims derived from data mining. Natural language processing could also help screen for relevant publications in full text databases such as Europe PMC.

Acknowledgements

Many thanks are due to the Wikibase.cloud user community and the Wikibase.cloud team at Wikimedia Deutschland (WMDE) for answering questions and support in troubleshooting; to Jorrit Poelen for help with GloBI integration and discussions on name alignment; and to WMDE for making Wikibase.cloud available in the open beta phase.

Data and code availability

Main URL for PPSDB: https://ppsdb.wikibase.cloud/

Export to tabular format for indexing in GloBI: https://github.com/kbseah/ppsdb-globi-export, archived on Zenodo: https://doi.org/10.5281/zenodo.12687626

Scripts for database maintenance: https://github.com/kbseah/ppsdb-utils, archived on Zenodo: https://doi.org/10.5281/zenodo.12805883

XML export of the entire PPSDB database on Internet Archive (16 Aug 2024): https://archive.org/details/wiki-ppsdbwikibasecloud_w

Figures and Tables

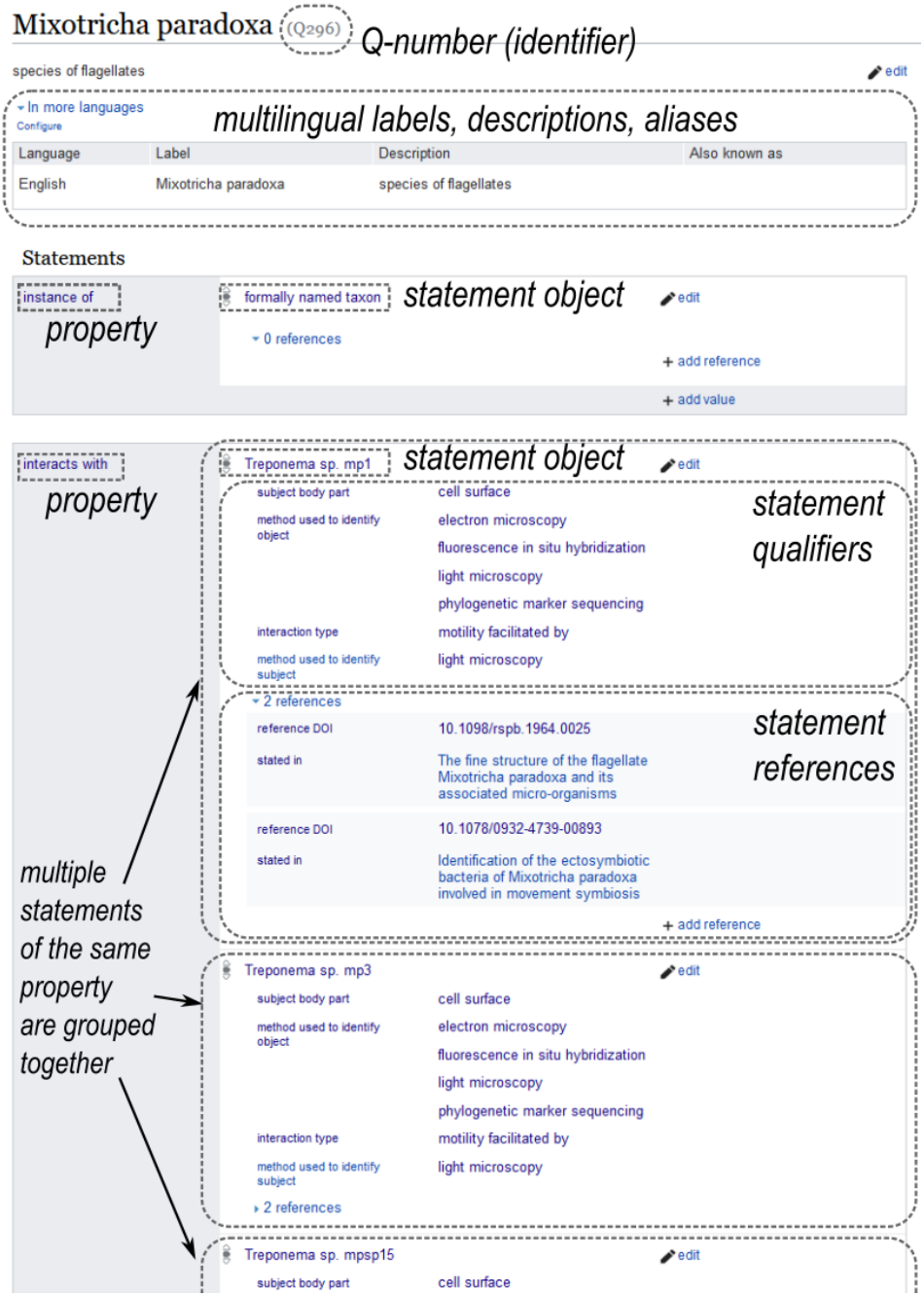

Figure 1. Annotated screenshot from the Wikibase item page for a host species, *Mixotricha paradoxa*, which illustrates how symbiotic interactions are represented as statements that link it to respective symbiont items, and which can be further qualified by additional information and references.

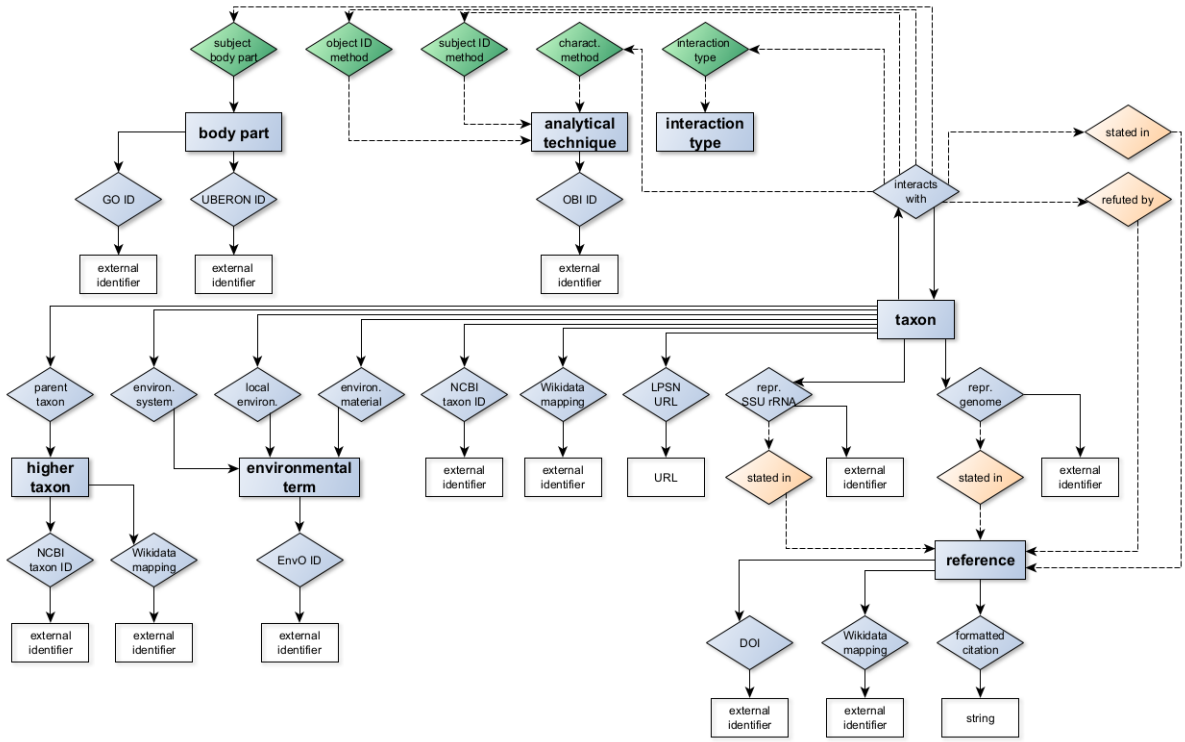

Figure 2. Relationships between items (blue rectangles), properties (diamonds), and other data types (white rectangles) in the PPSDB data model. Qualifier and reference relationships are depicted with dashed lines. Properties used in qualifiers or references are colored green and orange respectively. Subclasses of "taxon" and "environmental term" are not shown, for simplicity.

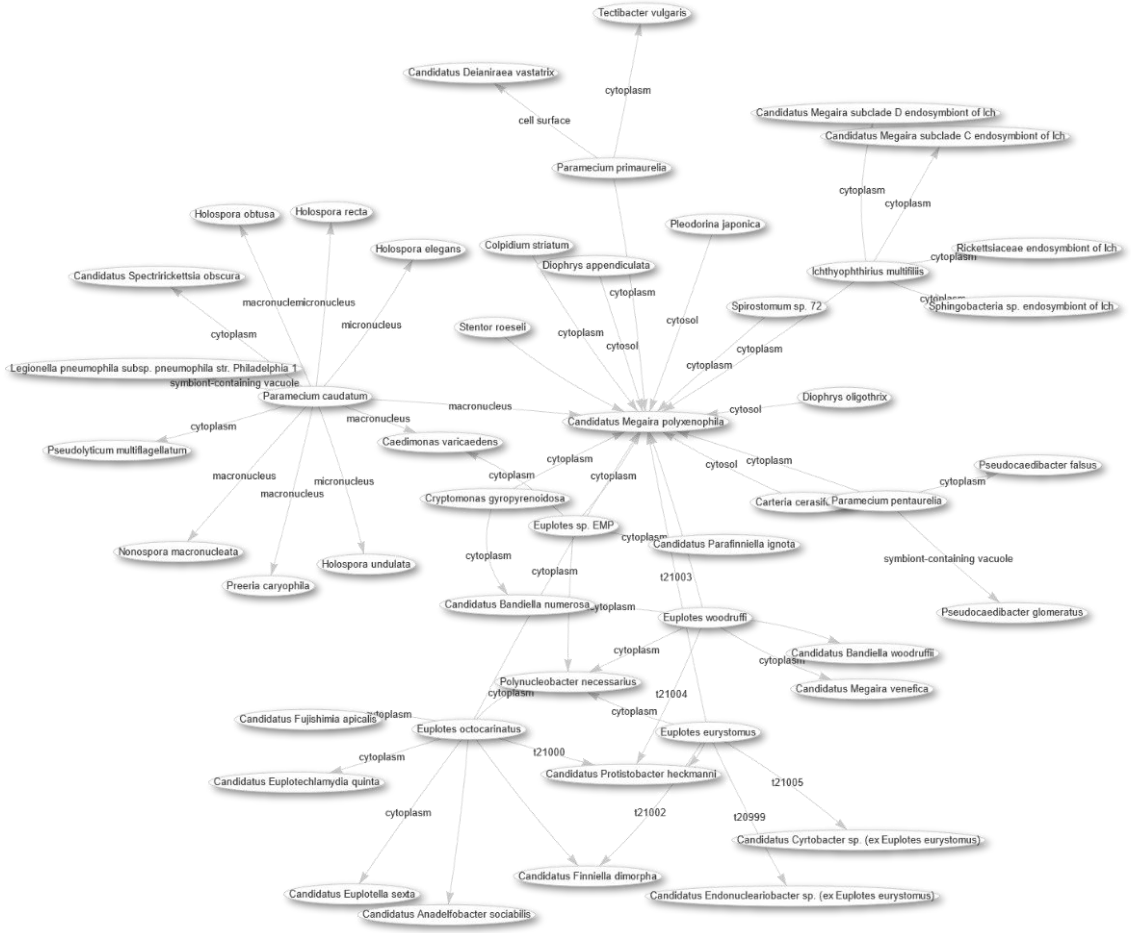

Figure 3. Graphical display of protist host species of the bacterial symbiont *Ca*. Megaira polyxenophila, and their other respective symbionts. The visualization was produced from a SPARQL query result by the Query Service engine included in Wikibase.cloud (https://tinyurl.com/2948dund).

Table 1. Concepts or entities represented in the database and relevant external databases, ontologies, or identifiers that they are linked to, if exact matches are available.

Concept/entity	**Relevant database, ontology, or identifier**
Taxonomy of the interacting organisms	NCBI Taxonomy Wikidata List of Prokaryotic Names with Standing in Nomenclature (LPSN)
Localization of symbionts in the host organism	Gene Ontology UBERON
Nature of the biotic interaction, if known or inferred	OBO Relations Ontology
Analytical methods used to identify organisms or determine the interaction type	OBI Evidence Ontology
Environment where the host organism was collected or isolated	Environment Ontology
Publication describing the symbiosis	DOI Wikidata
Nucleotide sequence records	Genbank accession

References

Adl S. M., Bass D., Lane C. E., Lukeš J., Schoch C. L., Smirnov A., Agatha S., Berney C., Brown M. W., Burki F., et al. 2019. Revisions to the classification, nomenclature, and diversity of eukaryotes. *J. Eukaryot. Microbiol.*, **66**:4–119.

Albrycht K., Rynkiewicz A. A., Harasymczuk M., Barylski J. & Zielezinski A. 2022. Daily Reports on Phage-Host Interactions. *Front. Microbiol.*, **13**:946070.

Ashburner M., Ball C. A., Blake J. A., Botstein D., Butler H., Cherry J. M., Davis A. P., Dolinski K., Dwight S. S., Eppig J. T., et al. 2000. Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**:25–29.

Ball G. H. 1969. Organisms living on and in protozoa. *In*: Research in Protozoology. Elsevier. p. 565–718.

Bandrowski A., Brinkman R., Brochhausen M., Brush M. H., Bug B., Chibucos M. C., Clancy K., Courtot M., Derom D., Dumontier M., et al. 2016. The ontology for biomedical investigations. *PLoS ONE*, **11**:e0154556.

Bauer F. & Kaltenböck M. 2012. Linked Open Data: The Essentials.

Bjorbækmo M. F. M., Evenstad A., Røsæg L. L., Krabberød A. K. & Logares R. 2020. The planktonic protist interactome: where do we stand after a century of research? *ISME J.*, **14**:544–559.

Boscaro V., Schrallhammer M., Benken K. A., Krenek S., Szokoli F., Berendonk T. U., Schweikert M., Verni F., Sabaneyeva E. V. & Petroni G. 2013. Rediscovering the genus Lyticum, multiflagellated symbionts of the order Rickettsiales. *Sci. Rep.*, **3**:3305.

Buttigieg P. L., Morrison N., Smith B., Mungall C. J., Lewis S. E. & ENVO Consortium. 2013. The environment ontology: contextualising biological and biomedical entities. *J. Biomed. Semantics*, **4**:43.

Buttigieg P. L., Pafilis E., Lewis S. E., Schildhauer M. P., Walls R. L. & Mungall C. J. 2016. The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *J. Biomed. Semantics*, **7**:57.

Chaudhri V. K., Baru C., Chittar N., Dong X. L., Genesereth M., Hendler J., Kalyanpur A., Lenat D. B., Sequeda J., Vrandečić D., et al. 2022. Knowledge graphs: Introduction, history, and perspectives. *AIMag*, **43**:17–29.

Davison H. R., Hurst G. D. D. & Siozios S. 2023. "Candidatus Megaira" are diverse symbionts of algae and ciliates with the potential for defensive symbiosis. *Microb. Genom.*, **9**.

Diefenbach D., Wilde M. D. & Alipio S. 2021. Wikibase as an Infrastructure for Knowledge Graphs: The EU Knowledge Graph. *In*: Hotho A., Blomqvist E., Dietze S., Fokoue A., Ding Y., Barnaghi P., Haller A., Dragoni M. & Alani H. (eds.), The semantic web – ISWC 2021. Vol. 12922. Lecture notes in computer science. Cham, Springer International Publishing. p. 631–647.

Fokin S. I. & Serra V. 2022. Bacterial symbiosis in ciliates (Alveolata, Ciliophora): Roads traveled and those still to be taken. *J. Eukaryot. Microbiol.*, **69**:e12886.

Gene Ontology Consortium, Aleksander S. A., Balhoff J., Carbon S., Cherry J. M., Drabkin H. J., Ebert D., Feuermann M., Gaudet P., Harris N. L., et al. 2023. The Gene Ontology knowledgebase in 2023. *Genetics*, **224**.

Hongoh Y., Sato T., Dolan M. F., Noda S., Ui S., Kudo T. & Ohkuma M. 2007. The motility symbiont of the termite gut flagellate Caduceia versatilis is a member of the "Synergistes" group. *Appl. Environ. Microbiol.*, **73**:6270–6276.

Huaman E., Huaman J. L. & Huaman W. 2023. Getting Quechua Closer to Final Users Through Knowledge Graphs. *In*: Lossio-Ventura J. A., Valverde-Rebaza J., Díaz E. & Alatrista-Salas H. (eds.), Information management and big data: 9th annual international conference, simbig 2022, lima, peru, november 16–18, 2022, proceedings. Vol. 1837. Communications in computer and information science. Cham, Springer Nature Switzerland. p. 61–69.

Husnik F., Tashyreva D., Boscaro V., George E. E., Lukeš J. & Keeling P. J. 2021. Bacterial and archaeal symbioses with protists. *Curr. Biol.*, **31**:R862–R877.

Koho M., Coladangelo L. P., Ransom L. & Emery D. 2023. Wikibase model for premodern manuscript metadata harmonization, linked data integration, and discovery. *J. Comput. Cult. Herit.*, **16**:1–25.

Kostygov A. Y., Karnkowska A., Votýpka J., Tashyreva D., Maciszewski K., Yurchenko V. & Lukeš J. 2021. Euglenozoa: taxonomy, diversity and ecology, symbioses and viruses. *Open Biol.*, **11**:200407.

La Scola B., Audic S., Robert C., Jungang L., de Lamballerie X., Drancourt M., Birtles R., Claverie J.-M. & Raoult D. 2003. A giant virus in amoebae. *Science*, **299**:2033.

Lamy-Besnier Q., Brancotte B., Ménager H. & Debarbieux L. 2021. Viral Host Range database, an online tool for recording, analyzing and disseminating virus-host interactions. *Bioinformatics*, **37**:2798–2801.

von Mering S., Gardiner L. M., Knapp S., Lindon H., Leachman S., Ulloa Ulloa C., Vincent S. & Vorontsova M. S. 2023. Creating a multi-linked dynamic dataset: a case study of plant genera named for women. *Biodivers. Data J.*, **11**:e114408.

Mihara T., Nishimura Y., Shimizu Y., Nishiyama H., Yoshikawa G., Uehara H., Hingamp P., Goto S. & Ogata H. 2016. Linking Virus Genomes with Host Taxonomy. *Viruses*, **8**:66.

Monteil C. L., Vallenet D., Menguy N., Benzerara K., Barbe V., Fouteau S., Cruaud C., Floriani M., Viollier E., Adryanczyk G., et al. 2019. Ectosymbiotic bacteria at the origin of magnetoreception in a marine protist. *Nat. Microbiol.*, **4**:1088–1095.

Mungall C. J., Torniai C., Gkoutos G. V., Lewis S. E. & Haendel M. A. 2012. Uberon, an integrative multi-species anatomy ontology. *Genome Biol.*, **13**:R5.

Nielsen F. Å., Mietchen D. & Willighagen E. 2017. Scholia, scientometrics and wikidata. *In*: Blomqvist E., Hose K., Paulheim H., Ławrynowicz A., Ciravegna F. & Hartig O. (eds.), The semantic web: ESWC 2017 satellite events. Vol. 10577. Lecture notes in computer science. Cham, Springer International Publishing. p. 237–259.

Pacheco A. R., Pauvert C., Kishore D. & Segrè D. 2022. Toward FAIR representations of microbial interactions. *mSystems*, **7**:e0065922.

Parks D. H., Chuvochina M., Chaumeil P.-A., Rinke C., Mussig A. J. & Hugenholtz P. 2020. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.*, **38**:1079–1086.

Parte A. C., Sardà Carbasse J., Meier-Kolthoff J. P., Reimer L. C. & Göker M. 2020. List of Prokaryotic names with Standing in Nomenclature (LPSN) moves to the DSMZ. *Int. J. Syst. Evol. Microbiol.*, **70**:5607–5612.

Pasulka A. L., Goffredi S. K., Tavormina P. L., Dawson K. S., Levin L. A., Rouse G. W. & Orphan V. J. 2017. Colonial Tube-Dwelling Ciliates Influence Methane Cycling and Microbial Diversity within Methane Seep Ecosystems. *Front. Mar. Sci.*, **3**.

Patterson D. J., Cooper J., Kirk P. M., Pyle R. L. & Remsen D. P. 2010. Names are key to the big new biology. *Trends Ecol. Evol.*, **25**:686–691.

Petroni G., Spring S., Schleifer K. H., Verni F. & Rosati G. 2000. Defensive extrusive ectosymbionts of Euplotidium (Ciliophora) that contain microtubule-like structures are bacteria related to Verrucomicrobia. *Proc Natl Acad Sci USA*, **97**:1813–1817.

Poelen J. H., Simons J. D. & Mungall C. J. 2014. Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecol. Inform.*, **24**:148–159.

Schoch C. L., Ciufo S., Domrachev M., Hotton C. L., Kannan S., Khovanskaya R., Leipe D., Mcveigh R., O'Neill K., Robbertse B., et al. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)*, **2020**.

Schrallhammer M., Castelli M. & Petroni G. 2018. Phylogenetic relationships among endosymbiotic R-body producer: Bacteria providing their host the killer trait. *Syst. Appl. Microbiol.*, **41**:213–220.

Schrallhammer M., Ferrantini F., Vannini C., Galati S., Schweikert M., Görtz H.-D., Verni F. & Petroni G. 2013. "Candidatus Megaira polyxenophila" gen. nov., sp. nov.: considerations on evolutionary history, host range and shift of early divergent rickettsiae. *PLoS ONE*, **8**:e72581.

Schulz F., Martijn J., Wascher F., Lagkouvardos I., Kostanjšek R., Ettema T. J. G. & Horn M. 2016. A Rickettsiales symbiont of amoebae with ancient features. *Environ. Microbiol.*, **18**:2326–2342.

Seah B. K. B. 2023. Paying it forward: Crowdsourcing the harmonisation and linking of taxon names and biodiversity identifiers. *Biodivers. Data J.*, **11**:e114076.

Shimizu C., Hitzler P., Gonzalez-Estrecha S., Goeke-Smith J., Rehberger D., Foley C. & Sheill A. 2023. The Wikibase Approach to the Enslaved.Org Hub Knowledge Graph. *In*: Payne T. R., Presutti V., Qi G., Poveda-Villalón M., Stoilos G., Hollink L., Kaoudi Z., Cheng G. & Li J. (eds.), The semantic web – ISWC 2023: 22nd international semantic web conference, athens, greece, november 6–10, 2023, proceedings, part II. Vol. 14266. Lecture notes in computer science. Cham, Springer Nature Switzerland. p. 419–434.

Shi Y., Queller D. C., Tian Y., Zhang S., Yan Q., He Zhili, He Zhenzhen, Wu C., Wang C. & Shu L. 2021. The Ecology and Evolution of Amoeba-Bacterium Interactions. *Appl. Environ. Microbiol.*, **87**.

Utami Y. D., Kuwahara H., Igai K., Murakami T., Sugaya K., Morikawa T., Nagura Y., Yuki M., Deevong P., Inoue T., et al. 2019. Genome analyses of uncultured TG2/ZB3 bacteria in "Margulisbacteria" specifically attached to ectosymbiotic spirochetes of protists in the termite gut. *ISME J.*, **13**:455–467.

Wardeh M., Risley C., McIntyre M. K., Setzkorn C. & Baylis M. 2015. Database of host-pathogen and related species interactions, and their global distribution. *Sci. Data*, **2**:150049.

Yilmaz P., Kottmann R., Field D., Knight R., Cole J. R., Amaral-Zettler L., Gilbert J. A., Karsch-Mizrachi I., Johnston A., Cochrane G., et al. 2011. Minimum information about a

548 marker gene sequence (MIMARKS) and minimum information about any (x) sequence
549 (MIxS) specifications. *Nat. Biotechnol.*, **29**:415–420.