Software Tool Article

GeneDB and Wikidata [version 1; peer review: 1 approved, 1 approved with reservations]

Magnus Manske, Ulrike Böhme, Christoph Pütche, Matt Berriman

Parasites and Microbes, Wellcome Trust Sanger Institute, Cambridge, CB10 1SA, UK

Abstract

Publishing authoritative genomic annotation data, keeping it up to date, linking it to related information, and allowing community annotation is difficult and hard to support with limited resources. Here, we show how importing GeneDB annotation data into Wikidata allows for leveraging existing resources, integrating volunteer and scientific communities, and enriching the original information.

Keywords

GeneDB, Wikidata, MediaWiki, Wikibase, genome, reference, annotation, curation

Corresponding author: Magnus Manske (mm6@sanger.ac.uk)

Author roles: Manske M: Conceptualization, Methodology, Software, Writing – Original Draft Preparation; Böhme U: Data Curation, Validation, Writing – Review & Editing; Pütche C: Project Administration, Writing – Review & Editing; Berriman M: Resources, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Wellcome Trust through a Biomedical Resources Grant to MB [108443].

Copyright: © 2019 Manske M et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Manske M, Böhme U, Pütche C and Berriman M. GeneDB and Wikidata [version 1; peer review: 1 approved, 1 approved with reservations] Wellcome Open Research 2019, 4:114 (https://doi.org/10.12688/wellcomeopenres.15355.1)

First published: 01 Aug 2019, 4:114 (https://doi.org/10.12688/wellcomeopenres.15355.1)
Introduction
The GeneDB website has presented genome annotation data from eukaryotic and prokaryotic pathogens\(^1\) sequenced by the Wellcome Sanger Institute for more than 15 years. The underlying data are stored in a database, designed using the Chado\(^2\) schema. The project was established to display genomes sequenced and annotated by the former Pathogen Sequencing Unit at the Sanger Institute, but over time the usage has changed. Now, genomes are stored and displayed if they are undergoing some level of curation or ongoing improvement. The site provides a way for curators and researchers to see changes to annotation long before those changes are integrated with other data types in a number of collaborating databases. To reflect the change of usage, where the website is often not the primary access point for many users, GeneDB has recently undergone a redesign and simplification\(^1\). In particular, the web-based genome annotation tool Apollo\(^1\) has been adopted as a major entry point for viewing genome data. While this delivers a structured, multi-track view of the genome and annotated genomic features (genes, ncRNAs, etc), the current version of Apollo has a limited capability for displaying the rich functional descriptions of individual genes that were a major feature of the previous GeneDB website.

Wikidata is a collaboratively edited, machine-readable and -writable knowledge base hosted by the Wikimedia Foundation, which also runs the collaboratively edited encyclopedia Wikipedia. Wikipedia has become the most accessed online encyclopedia and is unique in both its open, community-based editing, and a first port-of-call for public access to curated knowledge. Several bioinformatics projects make use of Wikipedia. The most successful of these is the Rfam project, where Wikipedia has been used to successfully manage free-text descriptions of RNA families\(^3\) for over a decade.

Wikidata currently contains 55 million items, which represent a superset of all Wikipedia article topics in over 300 languages, including biographical items, locations, species, artworks, scientific publications, etc. Amongst these items, Wikidata already stores human and mouse genes and proteins, as part of the Gene Wiki project\(^4\), which originally started on Wikipedia\(^4\).

Wikidata offers various application programming interfaces (APIs) to read or write information in an automated way, including a query service using SPARQL, a query language for data on the Semantic Web\(^5\). All these services are freely accessible by third-party users.

In the present study, we describe how we have exported the contents of GeneDB into Wikidata to ensure the long term sustainability of high value curated information and to make the annotated gene and protein information available to a wider audience. Within Wikidata, potentially anyone can contribute to the annotation, for instance by adding further external cross-references to third-party databases, linking gene and proteins to the scientific literature, or even short free-text descriptions. These community changes can be detected, checked, and, in appropriate cases, imported back into GeneDB.

We also describe utilising the Wikidata APIs to create a new version of the GeneDB website with content created solely based on Wikidata items. The design of the new GeneDB website closely mirrors the old one but now provides continuity and stability for incoming links from other websites. Furthermore, by building the site from Wikidata components, the new GeneDB website benefits from additional information and queries harvested from Wikidata.

Methods
Importing GeneDB into Wikidata
GeneDB exports its Chado database monthly into publicly accessible files (ftp://ftp.sanger.ac.uk/pub/genedb/releases/) in general feature format version 3 (GFF3) and gene association file (GAF) format.

These files are regularly parsed by bespoke code to create or update Wikidata items, for both genes and their protein products.

This includes the addition of GO terms, as well as the creation and usage of Wikidata items about the scientific publications containing the respective findings.

An item about a gene (example: https://www.wikidata.org/wiki/Q19044775) or a protein on Wikidata consists of labels and aliases, descriptions, and a list of statements. Each statement is comprised of the following: a property from a community-controlled vocabulary (e.g., “chromosome”, “found in species”, “GeneDB ID”); a value, usually a plain string or a link to another Wikidata item, but also a date, a location, or a number, depending on the property; an optional list of qualifiers; and an optional list of references.

When updating an item, elements are added, altered, or removed on Wikidata if the current GeneDB information is different, and GeneDB is the authoritative source. All other elements of the Wikidata item remain unchanged during updates. Updates are performed automatically, utilizing the publicly exported GFF and GAF files based on Chado.

Importing community changes from Wikidata into GeneDB
Community contributions on Wikidata can be divided into two parts. One part is the mass edit of items, by either bots (software-based robots that perform automated editing) or mass-editing tools. The other part are individual, usually manual edits, of low volume.

Only some edits are directly relevant to GeneDB; a new description of a protein in Dutch will not be imported into the Chado database, and can therefore be ignored. Likewise, the addition of external identifiers to sources not tracked by GeneDB can be ignored.

Individual edits that are both relevant to GeneDB, and not done by a Wikidata user on a “whitelist” of known, trusted users are summarized daily by an automated script, and sent to the GeneDB ticketing system for manual inspection. These...
changes are either ignored, reverted on Wikidata (e.g., vandalism), or imported into the GeneDB Chado database. The volume of such edits is quite low (~1/week) at this time, though we expect this to pick up with more members of the scientific community becoming aware of this venue into Wikidata.

**Implementation**

The code to import and update Wikidata items was written in Rust (rustc 1.36.0), using (amongst others) the rust-bio crate for GFF and GAF reading. Rust was chosen as a language for its speed, security, low resource consumption, and available crates (libraries) to build on. Some of these crates, such as MediaWiki and Wikibase (Wikidata) API handling, were (co-)developed by the corresponding author independently. See software availability for source code.

**Operation**

The Wikidata import code will run on any platform that Rust can be compiled for. Additional requirements are an internet connection, and login information for a Wikidata bot account.

The website operates client-side using JavaScript, and can be deployed on any standard web server. Besides Wikidata, no additional server-side support is required.

**Results/Use cases**

At this time (26/07/2019), 409,219 Wikidata items for genes (including 9,031 pseudogenes), and an additional 397,979 items for proteins, have been created (http://tinyurl.com/y6wpfym), covering 45 taxa.

By exporting from GeneDB (Chado) into Wikidata, the data are intricately integrated into its ecosystem creating new functions with minimal project-specific development; for instance, links to and between publications, which in turn link to authors, institutions, etc. These connections between items allow for complex queries that were not possible before (Figure 1).

**Gene pages**

To replicate the way genes were represented on the previous version of the GeneDB website, a pure HTML/JavaScript site using vue.js was created. JavaScript components are written as modules. All pages and components are designed to work on both desktop and mobile. Web pages for genes, proteins, species, chromosomes, GO term queries, and searches are generated on-the-fly utilising the Wikidata API and SPARQL interface, and Wikidata serves as the only back-end for these pages.

Each gene item, and its associated protein item(s), can be viewed on a separate page (Figure 2), that is rendered on-the-fly. This rendering includes a map of the gene on the chromosome, names, IDs, descriptions, links to other web resources (both from Wikidata statements, and auto-generated based on species and gene ID), a link to the Apollo browser view of the gene, a list of known orthologs.

Below that, each protein encoded by the gene is listed, as well as the known GO term ontology, complete with evidence and publication links, where available (Figure 3).

**Figure 1.** Part of the Scholia tool website about the PfCRT gene (https://tools.wmflabs.org/scholia/topic/Q19044775). These figures show connected topics, connected locations on a map, and the most prolific authors for this topic, respectively. All figures are generated live from Wikidata via its various APIs. All this information, and more, results from linking scientific publications to the genes in Wikidata.
Figure 2. Gene information on GeneDB (https://www.genedb.org/gene/PF3D7_07090000). All information comes exclusively from Wikidata.

If Wikidata contains items about publications that have the gene or protein as a “main subject”, these publications are listed at the bottom of the page. This is an example of additional, on-topic information that Wikidata provides on top of the GeneDB dataset.

Gene/protein pages link out to other web-based databases via Wikidata-stored (e.g., UniProt) or computed (e.g., PlasmoDB) URLs. GO terms show supporting information, including citations of, and links to, the original publications. Also, a list of all publications on Wikidata with the respective gene as a subject.
### Ontology

<table>
<thead>
<tr>
<th>Cell component</th>
<th>Determination method</th>
<th>Added to CHADO</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>nuclear condensin complex</strong></td>
<td>InterPro ID: IPR012371</td>
<td>2019-03-10</td>
<td>GOREF: 0000002</td>
</tr>
</tbody>
</table>

**Figure 3.** GO term ontology ([https://www.genedb.org/gene/PF3D7_1135600](https://www.genedb.org/gene/PF3D7_1135600)), with sources.

is available on both GeneDB (example: [https://www.genedb.org/gene/PF3D7_1200600](https://www.genedb.org/gene/PF3D7_1200600)) and the Scholia tool (example: [https://tools.wmflabs.org/scholia/topic/Q18971176](https://tools.wmflabs.org/scholia/topic/Q18971176)).

**Other functionality**

The GeneDB search function utilizes Wikidata search, letting users find genes by name, alias, ID, and related information, across all covered species. The search will only return genes on Wikidata with a GeneDB identifier.

Each species in GeneDB has its own page (Figure 4), showing basic information about the species, links to other web resources, to the Apollo browser, and a list of chromosomes linking to the genes located there.
Genes/proteins annotated with specific GO terms can be listed, grouped by species (Figure 5). These lists are linked from every GO term on a gene/protein page.

**Discussion**

Wikidata has provided GeneDB with a venue to host and publish its data, and to invite community edits, without giving up on GeneDB’s curation authority. The simplified maintenance of the HTML/JS-only GeneDB website, compared to a previous one that combined a frontend and backend solution, frees technical and personnel resources. Linking genes and proteins to, and from, other Wikidata items allows for novel methods of querying the data, and for new questions to be asked. Publishing on Wikidata also exposes the data in new ways and potentially engages with a broader public audience.

---

**Figure 4.** Species page for *Plasmodium gallinaceum*, using information and image from Wikidata.
Data availability

Source data
Chado-based GFF and GAF files
ftp://ftp.sanger.ac.uk/pub/genedb/releases/

Accession numbers in use cases
GeneDB IDs
Pathogen genomic data from GeneDB, Accession number PF3D7_0709000. https://identifiers.org/genedb:PF3D7_0709000


Plasmodium gallinaceum data from Wikidata, Accession number: Q7201888. https://identifiers.org/wikidata:Q7201888

GO terms

Underlying data
All data underlying the results are available as part of the article and no additional source data are required.

Software availability
GeneDBot, the code that updates Wikidata from Chado:

Source code: https://github.com/sanger-pathogens/genedbot
Archived source code at time of publication: http://doi.org/10.5281/zenodo.335200
License: GPL v3.0
The HTML/JS code for the GeneDB website (Wikidata display code is mostly in the htdocs/wd directory):

Source code: https://github.com/sanger-pathogens/GeneDB-static

Archived source code at time of publication: http://doi.org/10.5281/zenodo.3352003

License: GPL v3.0

Grant information
This work was supported by the Wellcome Trust through a Biomedical Resources Grant to MB [108443].

References


Open Peer Review

Current Peer Review Status:  ✔️  📑

Version 1

Reviewer Report 27 August 2019

https://doi.org/10.21956/wellcomeopenres.16772.r36100

© 2019 Waagmeester A. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Andra Waagmeester
Miciel, Antwerp, Belgium

This paper describes work done in linking data from the GeneDB website in Wikidata. GeneDB captures genome annotation data. The article also describes how a feedback loop is created enriching GeneDB with content from Wikidata. This is an exciting development where the UI directly leverages content from Wikidata, instead of using in-house caching databases.

I have some minor comments.

1. Wikidata uses a CC0 license on captured data. It is not clear if this is compatible with the applicable Guidelines on the use of data in publications on GeneDB (https://www.sanger.ac.uk/legal/). The terms for usage require explicitly attribution, which CC0 does not. I am curious how the authors solved this. A section on legalities of sharing GeneDB on Wikidata would be a welcome addition.

2. “The most successful of these is the Rfam project, where Wikipedia has been used to successfully manage free-text descriptions of RNA families for over a decade.” Why is this the most successful project, who are the second and third runner up?

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly
Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Biomedical informatics, Wikidata, Semantic Web,

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

- Sebastian Burgstaller-Muehlbacher
  Max Perutz Labs, Vienna Biocenter, Medical University of Vienna, Vienna, Austria

The authors describe a new implementation and technical foundation of GeneDB, a database for eukaryotic and prokaryotic parasites. As a new technical foundation, the authors use Wikidata, an open, semantic web enabled graph database operated by the Wikimedia Foundation. To integrate GeneDB data into Wikidata, the authors wrote bots (robots/database clients) using the language Rust and imported GeneDB genes, proteins and Gene Ontology annotations into Wikidata. Furthermore, they created a website which renders the data imported into Wikidata in a comprehensive and more user-friendly way than a raw Wikidata data item would. Also, they integrate genome track views using Apollo.

Overall, the study opens up a new dataset to the public domain entirely. Data are imported and made available via the Wikidata API and the Wikidata SPARQL endpoint, allowing users of the data powerful queries. Having these data available and easily accessible to the parasite research community worldwide seems crucial. Easy access is also enabled by the web interface the authors implemented. Moreover, due to the nature of Wikidata, contributions by the (research) community are relatively easy to accomplish. The authors also provide a reasonable approach for feeding back (research) community contributions into GeneDB, while avoiding vandalism.

Integration with the Apollo genome browser also provides a genome view, typical for a gene/genome centered database. However, the authors should describe in some more detail how they setup Apollo and where Apollo gets its genome tracks from (e.g. DNA sequences). This is relevant as such data typically cannot be stored in Wikidata and thus need to be drawn from a different source.

The manuscript is well-written, however the introduction should also mention the work of Putman *et al.* (2017), who follow a similar technical approach but for a different set of genomes. The discussion is rather brief and could profit from addressing the authors long-term maintenance plans, especially regarding curation of the inflow of community contributions to GeneDB.
In summary, this work by Manske et al. is very well performed, except for the minor points mentioned above. It’s an important data release for the parasite research community and enables, for the first time, collaborative work on genomic data integration and data annotation within this community. This should lower greatly the hurdles for worldwide collaboration on alleviating the severe diseases associated with many of the parasites in GeneDB.

Ad citation 7 of the manuscript: This preprint has now been published as Burgstaller-Muehlbacher et al. (2016).

References

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, semantic web, knowledge networks, deep learning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.