

1 WikiGenomes: an open Web application for community consumption and curation of gene
2 annotation data in Wikidata.

3
4 Tim E. Putman¹, Sebastien Lelong¹, Sebastian Burgstaller-Muehlbacher¹, Andra Waagmeester²,
5 Colin Diesh⁴, Nathan Dunn³, Monica Munoz-Torres³, Gregory S. Stupp¹, Andrew I. Su¹ and
6 Benjamin M. Good¹

- 7
8 1. Department of Molecular and Experimental Medicine, The Scripps Research Institute, La
9 Jolla, USA
10 2. Micelio, Antwerp, Belgium
11 3. Environmental Genomics and Systems Biology Division, Lawrence Berkeley National
12 Laboratory, Berkeley, California.
13 4. Division of Animal Sciences, University of Missouri, Columbia, MO 65211, USA

14
15
16 {tputman, bgood, asu}@scripps.edu

17 **Abstract**

18 With the advancement of genome sequencing technologies, new genomes are being
19 sequenced daily. While these sequences are deposited in publicly available data warehouses,
20 their functional and genomic annotations (beyond genes which are predicted automatically)
21 mostly reside in the text of primary publications. Professional curators are hard at work
22 extracting those annotations from the literature for the most studied organisms and depositing
23 them in structured databases. However, the resources don't exist to fund the comprehensive
24 curation of the thousands of newly sequenced organisms in this manner. Here, we describe
25 WikiGenomes (wikigenomes.org), a web application that facilitates the consumption and
26 curation of genomic data by the entire scientific community. WikiGenomes is based on
27 Wikidata, an openly editable knowledge graph with the goal of aggregating published
28 knowledge into a free and open database. WikiGenomes empowers the individual genomic
29 researcher to contribute their expertise to the curation effort and integrates the knowledge into
30 Wikidata, enabling it to be accessed by anyone without restriction.

31

32 Introduction

33 Sequencing an organism's genome has become a routine procedure in the life sciences, but
34 until that genome is annotated, it tells us very little about the organism. The knowledge captured
35 in genomic and functional annotations provides the 'blueprint' of the biology of the organism that
36 can be leveraged to drive all manner of scientific inquiry. The knowledge represented by
37 annotations is mostly published in the free text of scientific journal articles. In order to make this
38 type of knowledge computable and accessible as structured annotations, we typically rely on
39 centralized curation efforts like those supported by model organism databases such as the
40 Zebrafish Model Organism database (ZFIN) (1) and Mouse Genome Informatics (MGI) (2).
41 Unfortunately, the vast majority of sequenced genomes do not have a dedicated curation team,
42 hence new curation models need to be explored.

43
44 Wikidata is a recently developed project of the Wikimedia Foundation that enables
45 the collaborative construction of a centralized graph database (3). It was initially created as a
46 central hub for structured data across all the language-specific Wikipedias, but its potential
47 applications are much broader. Wikidata follows the Resource Description Framework (RDF)
48 (<https://www.w3.org/RDF/> (4)) model, the W3C standard for data interchange. RDF links related
49 entities together into 'triples', defined by a subject concept, an object concept, and a predicate
50 that describes the relationship between them. This semantic structure allows the modeling of
51 complex systems in a queryable knowledge graph (3,5). The contents of Wikidata currently
52 represent 28 million concepts, from all domains of knowledge, spanning over 1.3 billion triples
53 (grafana.wikimedia.org/dashboard/db/wikidata-query-service).

54
55 While other resources have also mapped their data onto an RDF data model (6–8), Wikidata is
56 unique in that it allows both read and write access. Wikidata content can be both queried and
57 edited programmatically via an application programming interface (API)
58 (www.wikidata.org/w/api.php) and can be queried using the Wikidata Query Service (WDQS)
59 (query.wikidata.org) via the SPARQL Protocol and RDF Query Language (4,9,10). SPARQL is a
60 query language for graph databases similar in syntax to SQL for relational databases.

61
62 Notably, all Wikidata content is in the public domain, eliminating the potential for problematic
63 restrictions on downstream reuse and redistribution (11). It is populated through various
64 programs ("bots") (12) en masse, and through the discrete contributions of individual users
65 through Wikidata's own web interface or through a handful of tools developed by Wikimedia
66 Foundation's WMF Labs (tools.wmflabs.org/).

67
68 There is growing appreciation within the bioinformatics community for the potential of Wikidata
69 as an open resource that can be populated, queried and edited by anyone. However, Wikidata
70 has not yet widely reached the biological domain experts who have much to gain and
71 contribute. Taking advantage of the querying capabilities of Wikidata can be a challenge for
72 biologists as structured query languages like SPARQL are not commonly part of a researcher's
73 toolkit. Further, the domain-specific data models (e.g. the patterns of connections between
74 genes, proteins, and their annotations
75 (www.wikidata.org/wiki/Wikidata:WikiProject_Molecular_biology)) underlying the content in the
76 Wikidata graph are not readily apparent nor automatically enforced by the generic Wikidata web
77 interface. This makes providing meaningful contributions to the graph a challenge for the
78 researcher and a roadblock for community curation in general. While Wikidata offers a powerful
79 technical platform for constructing an open community knowledge base of unprecedented
80 scope, it cannot reach this potential without the large-scale participation of experts in the
81 domains of knowledge that it seeks to represent. To attract specialists, e.g. life scientists, its

82 content must be presented in compelling applications that are more useful than the applications
83 that these specialists currently use to do their work.

84
85 Here, we describe WikiGenomes (www.wikigenomes.org), the first domain-specific application
86 built on Wikidata, tailored to the needs of biomedical researchers. WikiGenomes is a Web
87 application that is designed and built to allow the genomic researcher to contribute to curating
88 the knowledge they are discovering. It facilitates both the consumption and community curation
89 of genomic data in Wikidata, extending the reach of the biocuration effort deeper into the long
90 tail of sequenced genomes.

91 **Results**

92 *Wikidata*

93 Our group and others have been working to populate Wikidata with a comprehensive and
94 centralized knowledge graph that represents biomedical knowledge in a structured, queryable,
95 and computable format (5,12). Previously, we loaded genomic data from a variety of organisms
96 including *Homo sapiens* “human” (www.wikidata.org/wiki/Q15978631), *Mus musculus* “mouse”
97 (www.wikidata.org/wiki/Q83310), *Rattus norvegicus* “brown rat”
98 (www.wikidata.org/wiki/Q184224), *Saccharomyces cerevisiae* S288c “baker’s yeast”
99 (www.wikidata.org/wiki/Q27510868), and *Macaca nemestrina* “Southern pig-tailed macaque”
100 (www.wikidata.org/wiki/Q618026). In this work, we systematically expanded this list to include
101 all genes (390,719) and proteins (372,178) from the 120 NCBI prokaryotic reference genomes
102 (<https://www.ncbi.nlm.nih.gov/genome/browse/reference/>) (**Figure 1**). These organisms are all
103 reasonably well-studied, but in general did not have community-maintained genome databases.
104 In addition to organism/genetic data, we contribute to and maintain data for proteins, chemical
105 compounds and diseases (5,12,13).

106
107 To ensure that the work we are doing in Wikidata is trusted as high quality by the scientific
108 community, our group references the source and provenance of all claims in a consistent and
109 structured way. Direct links to the data sources allow the user to see for themselves the source
110 of an annotation and read further in the referenced publication to come to their own
111 conclusions. Standards for representing the evidence and provenance of any claim made using
112 our infrastructure is outlined in a detailed project wiki
113 (<https://www.wikidata.org/wiki/User:ProteinBoxBot/evidence>). References can be accessed
114 either manually by the web interface or programmatically via the Wikidata API
115 (<https://www.wikidata.org/w/api.php>).

116
117 In addition to recording data provenance, Wikidata claims can be qualified by how the initial
118 claim was made. For example a Gene Ontology (GO) annotation can be qualified using the
119 Gene Ontology Evidence Codes to indicate the determination method for this claim. The
120 combination of references and qualifiers provides the necessary context to the stated claim,
121 making it more trustworthy and useful. This model allows the researcher to, for example, treat
122 electronically curated annotations differently from those that were manually curated.

123 124 *Data Aggregation and Maintenance in Wikidata*

125 As a foundation, we loaded key information for each genome about each gene and gene
126 product into Wikidata. Wikidata is not simply an open database into which data can be
127 dumped. It is a structured knowledge graph that requires careful consideration and community
128 consensus in its design. Currently, these entity models in Wikidata are not comprehensive
129 mirrors of those models in their sources (e.g. NCBI, UniProt, etc...). We have chosen to

130 aggregate key data from a variety of sources, creating a minimal viable model that can be
131 expanded on by us or others. The basic data in Wikidata for all the organisms we have loaded
132 includes genomic position and orientation, Gene Ontology annotations, Enzyme Commission
133 Numbers, InterPro domains and external database identifiers. The data model that we have
134 created in Wikidata, and the sources of the annotations and entities are illustrated and reported
135 in Figure 1.

136
137 <FIGURE 1 HERE>

138
139 Data retrieval, standardization, loading and maintenance are done routinely (currently monthly)
140 via our publically available WikidataIntegrator Python package
141 (<https://github.com/SuLab/WikidataIntegrator> and <https://pypi.python.org/pypi/wikidataintegrator>)
142 as well as M, and we are working towards custom update schedules that would follow each
143 source's update schedule. WikidataIntegrator interacts with the Wikidata API to create new or
144 edit existing Wikidata items when appropriate. The data points are mapped to the appropriate
145 "properties" in Wikidata (**Supplemental Table 1**). Mapping these properties to a new semantic
146 model in Wikidata is key to aggregation of data from different sources into a single cohesive
147 graph that represents our collective knowledge of each concept and the relationships between
148 those concepts.

149
150 An example of other data we have loaded and maintain is the Gene Ontology
151 (<http://www.geneontology.org/>), which we have loaded to Wikidata to provide semantic
152 framework for gene product annotations. The structure was represented as a hierarchical tree
153 using the Wikidata 'Subclass of' (P279) property. We programmatically gathered professionally
154 curated Gene Ontology annotations from the European Bioinformatics Institute's QuickGo API
155 (<https://www.ebi.ac.uk/QuickGO/>), and NCBI's Gene Database FTP resources
156 (<ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz>) for all of the annotated gene products in Wikidata
157 (5,12,13). Any gene product item in Wikidata that is annotated with a specific GO term will be
158 linked to that GO term's Wikidata item via the appropriate Wikidata GO property (e.g. *molecular*
159 *function P680*) (**Figure 1**). Our methods for loading and maintaining other domains, such as
160 disease and chemical data, are described in our previous works (12,13).

161
162 *WikiGenomes*

163 To allow non-developer biologists to interface with the genomic data knowledge graph that we
164 developed in Wikidata, we built WikiGenomes. WikiGenomes currently supports the 120 NCBI
165 Bacterial reference genomes (<https://www.ncbi.nlm.nih.gov/genome/browse/reference/>) and
166 their genes and gene products. We designed WikiGenomes to allow a user to consume
167 genomic data from any organism in Wikidata in a single web application. The user begins by
168 selecting an organism using a type-ahead search feature on the application's landing
169 page. Once an organism has been selected, the application is directed to the primary interface
170 page where the data for the selected organism is loaded and displayed. In addition to providing
171 access to genetic data for these organisms, WikiGenomes also allows users to directly
172 contribute genomic and protein annotations to the Wikidata knowledge graph (**Figure 2**).

173
174 <FIGURE 2 HERE >

175
176 The primary page of the interface is divided into four windows that render the data for the
177 selected organism: its genome, genes/proteins, and annotations for the currently displayed
178 gene/protein. To select and load data, there are two forms that allow organism and gene/protein
179 selection (**Figure 3**). Each of the data elements are drawn dynamically from the Wikidata

180 knowledge graph using the WDQS SPARQL query system. The genome data is rendered in the
181 open source JBrowse genome browser (14,15). The WebApollo and JBrowse development
182 teams expanded on already existing code in the JBrowse framework to allow JBrowse to load
183 annotations gathered from Wikidata SPARQL queries to the WDQS endpoint (16). The DNA
184 sequences themselves are curated by the RefSeq Project (16) and because bulk primary data is
185 not suitable for storing in Wikidata, are collected monthly from GenBank's FTP web service
186 (See **Supplementary File 1** for all Wikidata queries that drive the system and **Supplementary**
187 **Table 2** for the FTP paths of data collected from NCBI). The open source code for
188 WikiGenomes is available at <https://github.com/putmantime/CMOD.Django>.

189
190 <FIGURE 3 HERE>

191
192 While bioinformaticians can programmatically gather data and upload it to Wikidata,
193 WikiGenomes provides a domain specific editing interface for non-programmers. WikiGenomes
194 distills editing tasks into straightforward forms that guide the user through the process of
195 building annotations. Internally, the server translates the data entered through these forms into
196 the structures needed to populate the Wikidata knowledge graph and submits them via the
197 Wikidata API (**Figure 2**). WikiGenomes uses the WikiMedia OAuth extension to allow users to
198 edit Wikidata using their own Wikidata account (www.mediawiki.org/wiki/Extension:OAuth). This
199 makes it possible to utilize the Wikidata edit history to track the contributions of individual
200 editors, potentially offering ways to reward good editors as well as mechanisms for detecting
201 vandalism automatically ([Good 2012](#)).

202
203 *Community Curation Forms*

204 There are currently two editing forms in WikiGenomes, one for adding functional annotations
205 and one for adding operon annotations. The forms are designed to allow a user to curate data
206 from a publication by prompting the user to provide the necessary information and references to
207 make a useful contribution to Wikidata (**Figure 4**).

208
209 <FIGURE 4 HERE>

210
211 *Gene Ontology Form*

212 Molecular function, cellular component and biological process annotations types are added to
213 Wikidata (and subsequently WikiGenomes) through the Gene Ontology Annotation Form
214 (**Figure 4A**). The WikiGenomes GO form allows users to make structured GO annotations to
215 gene products by identifying an annotation from a primary publication, mapping the annotation
216 to the proper GO term, and then pushing it to Wikidata referencing the publication. We have
217 chosen to adhere to the GO annotation model, rather than user generated free-text, to guide the
218 annotation process in a consistent manner. This is accomplished by using the form to search
219 Wikidata by typeahead for the proper GO term (as mentioned we maintain the GO ontology in
220 Wikidata), selecting the proper GO evidence code ([http://geneontology.org/page/guide-go-
evidence-codes](http://geneontology.org/page/guide-go-evidence-codes)) to provide a determination method (i.e. EXP: Inferred from Experiment
222 <http://geneontology.org/page/exp-inferred-experiment>), and select the publication as a reference
223 using the PubMed Identifier (PMID) (**Figure 4A**). When the annotation is submitted, the
224 annotation data is pushed to Wikidata using the Wikidata API and the annotation will be
225 viewable in WikiGenomes shortly after (there is ~5 min delay as the Wikidata SPARQL endpoint
226 updates).

227
228 *Operon Form*

229 Regulatory annotations are rare in large data warehouses like NCBI and regulatory annotation
230 curation would greatly benefit from community involvement. To this end we created a form to
231 allow user curation of prokaryotic operons (**Figure 4B**). The WikiGenomes Operon Form works
232 similarly to the GO form in using typeahead search boxes to retrieve relevant data from
233 Wikidata, as well as input boxes to upload novel data. The user can create a new operon or
234 add genes to an existing operon, input the determination method and reference the PMID, then
235 submit that annotation and (as in the GO Form) our backend framework will do the work of
236 creating the structured and standardized annotation in Wikidata.

237

238 Discussion

239 WikiGenomes was built as a proof a concept that Wikidata can serve as a universal database
240 backend for domain specific applications. WikiGenomes is in the stage of functional prototype
241 and usage stats are thus not yet available. WikiGenomes provides a working demonstration of
242 a new technical pattern that, we posit, offers the potential to fundamentally change how
243 biocuration is organized in ways that, over time, will result in a more efficient movement of
244 knowledge throughout the scientific community. While this is not yet a tool in common usage,
245 we believe that it shows the potential of centralized, community owned knowledge aggregation
246 projects like Wikidata, and hope it will ultimately help stimulate community curation at both the
247 level of individual researchers and bioinformatics laboratories.

248 *Distributed Curation and Access, Centralized and Integrated Content*

249 The incorporation of operon information for *Chlamydia pneumoniae* into WikiGenomes
250 illustrates the pattern of centralizing content while decentralizing control. Despite prokaryotic
251 operons being heavily studied, the actual annotations are rarely included in genomic assemblies
252 and often reside in supplemental tables of primary publications. *Chlamydia pneumoniae* is a
253 pathogen affecting animals and humans causing lung infections
254 (www.cdc.gov/pneumonia/atypical/cpneumoniae/index.html). The operons of the C.
255 *pneumoniae* strain CWL029 genome were annotated in a 2011 study (17) but access to this
256 information was only provided in the form of a supplementary Excel file. To our knowledge,
257 none of those ~200 experimentally derived annotations have been curated in GenBank or in
258 operon-specific databases such as DOOR (18) or ODB3 (19). Illustrating the 'small data to big
259 data' approach of WikiGenomes, we loaded this content programmatically via the Wikidata API.
260 Individual domain experts could also have loaded these data via the operon form (**Figure 4b**).
261 Using either method, these annotations are now fully-linked data, viewable as a track in the
262 WikiGenomes browser and in any other application drawing content from the open Wikidata
263 knowledge base. Integrating this knowledge into the central Wikidata knowledge graph creates
264 queryable connections to other organisms, diseases, drugs and small molecules. The more
265 extensive the knowledge graph becomes, the more it can be used to explore biological research
266 questions (5),(12,13).

267 *Stepping Towards Community Curation*

268 So far, the great majority of content accessible through WikiGenomes has been added by our
269 research group. While this content is valuable, it is only a starting point on the path towards an
270 open, community-curated knowledge graph for biology. The fundamental logic of the
271 community curation pattern has been well-described many times (20–22). By incorporating the
272 scientific community into the process of curating the knowledge that they are generating and
273 consuming, we could, in theory, dramatically increase the scale at which biocuration can
274 collectively operate. Here, as usual, the devil is in the details. With certain notable exceptions,
275 most attempts to stimulate community curation fail to attract enough editors to achieve their

276 goals. With the WikiGenomes application, we make no claim that we have solved this challenge
277 of incentives and recruitment. To realize its vision, WikiGenomes will need to transition to a
278 model where the content consumers become the content producers. To make this transition, a
279 number of problems must be addressed. In facing them, some inspiration can be taken from
280 one of WikiGenomes' most important and successful predecessors, the Gene Wiki effort to
281 organize functional gene information in the context of Wikipedia articles (23).

282
283 The key distinction between the Gene Wiki and other projects with very similar goals but less
284 success in recruiting contributors (20,24), was that it was embedded directly in Wikipedia. The
285 Wikipedia context provided the Gene Wiki with immediate discoverability (e.g. Gene Wiki
286 articles rank highly in Google search results), connection to a large community of editors, and a
287 proven social and technical framework for supporting large-scale community content
288 creation. Wikidata now offers most of these same characteristics for structured information. Its
289 content is highly discoverable, with hyperlinks coming in from nearly every Wikipedia article and
290 a high performance query engine that provides a single point of entry into all of its data. It has
291 already attracted thousands of editors from the Wikipedia communities as well as external
292 groups more focused on structured data. It offers a user interface for human editing of content
293 as well as an effective API for programmatic updates. Finally, communities are forming around
294 it in similar ways to Wikipedia. There are social structures in place for administration of the
295 higher-level data structures it utilizes (e.g. the allowed properties) and emerging domain-
296 focused communities such as WikiProject Medicine and WikiProject Molecular and Cellular
297 Biology that work to build and maintain domain-specific content within the larger knowledge
298 graph.

299
300 With all of those similarities, one key distinction is that structured information is much more
301 difficult to edit and to present effectively than the hypertext of Wikipedia. While the
302 Wikidata interface (www.wikidata.org) provides a useful, general-purpose foundation for
303 navigating and editing its content, it is not nearly sufficient as a way to connect end-users with
304 the information they need. WikiGenomes provides a genome-focused, editable view of
305 biological knowledge in Wikidata; it offers the large community of scientists an easy way to both
306 access and share knowledge in a structured way, thus opening a door towards open community
307 curation.

308 309 *Conclusions and Next Steps*

310 While tools like WikiGenomes are a step in the right direction for converting 'small data' to 'big
311 data' several key problems remain to be solved in the community curation model. Interfaces
312 like WikiGenomes that provide access to content drawn from openly editable public sources
313 should consider ways to help users assess the trustworthiness of content based on automatic
314 inspection of the provenance information provided in associated references as well as in the
315 histories of the editors who generated the content. In addition to making trust computable, such
316 provenance information will also be of value in attributing the work of the curators who devote
317 time to these open resources - potentially providing an aspect of the solution to the incentive
318 problem. Aside from presenting trust-related information, the forms for adding content should
319 be designed in ways that help domain experts generate high-quality annotations in the first
320 place. The addition of documentation, logical constraints based on underlying ontologies, and
321 helper interfaces are areas where the experience of the existing biocuration resources would be
322 very valuable.

323
324 WikiGenomes is a first foray into the world of applications that could be built with the content
325 that is growing in Wikidata. As that content expands and diversifies, many new applications will

326 need to be created to bring that information most effectively to the people that need it and who
327 can contribute to it. Wikidata is unlikely to be the end-all and be-all of open biological
328 knowledge bases. While it has a number of advantages that make it an ideal platform for
329 initiating the endeavor of linking all biological knowledge, it has limitations in terms of content
330 scope that will eventually need to be overcome through the provision of equally open systems
331 that are tailored more specifically to biological content.

332
333 Immediate further development of WikiGenomes is focused on teaming up with biocuration
334 groups to help refine the user experience and identify important areas where the data model
335 needs to be expanded. One example would be to work further with the WebApollo team to
336 integrate WebApollo [\(25\)](#) into WikiGenomes, taking advantage of WebApollo's annotation
337 features. Another useful feature we will work towards implementing is the ability to export data
338 from Wikidata via WikiGenomes in standard formats (e.g. GFF3). We hope to eventually create
339 a tool in WikiGenomes that brings the biocuration and basic research communities together,
340 creating a more efficient system that will allow for more comprehensive reference genome
341 curation of biological knowledge than is currently possible. Eventually we would make
342 WikiGenomes a deployable toolkit that smaller research communities can customize and deploy
343 while easily sharing their work with the global community.

344

345

346 **Acknowledgements**

347 We would like to thank the Wikimedia foundation for supporting the Wikidata project, and
348 especially the Wikidata community for their efforts in populating it. This project was supported
349 by the National Institutes of General Medical Sciences (R01GM089820) and NIH Common Fund
350 programs for Big Data to Knowledge (U54GM114833) and Extracellular RNA Communication
351 (U54DA036134).

352

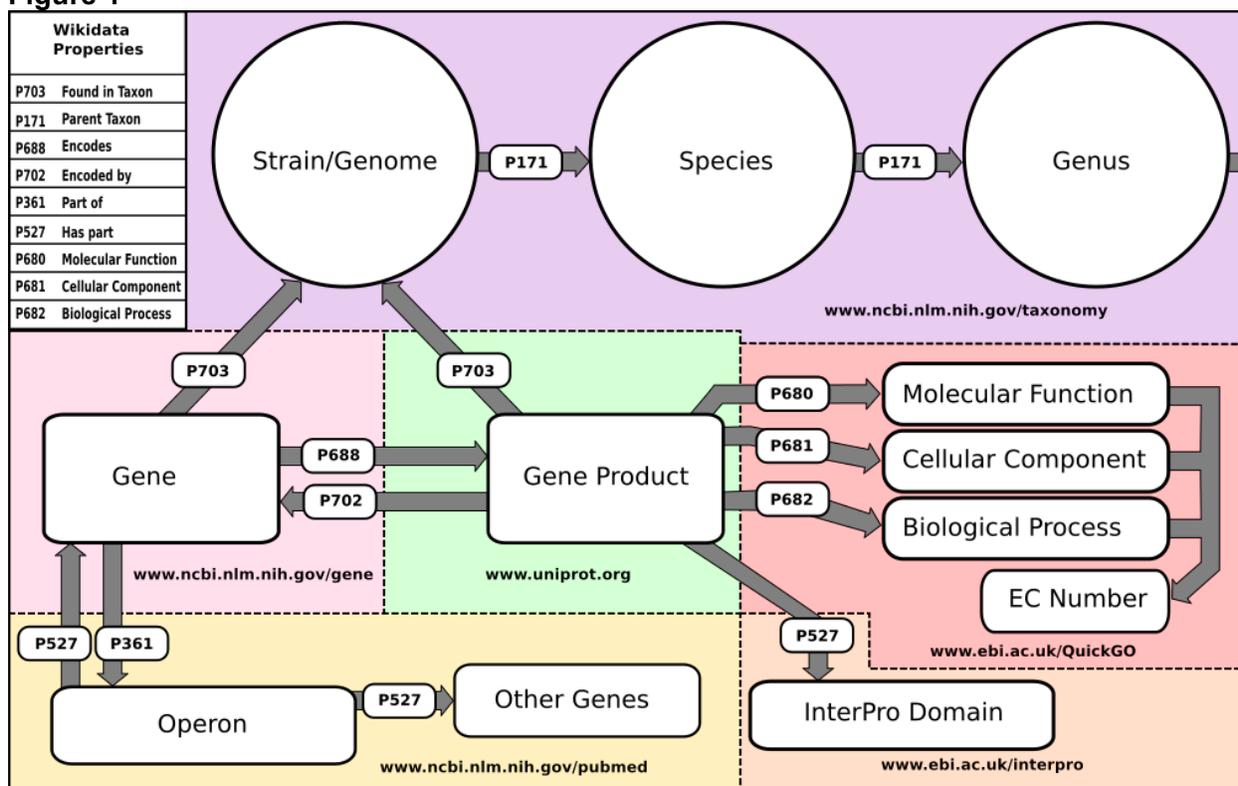
353 References

- 354 1. Howe, D. G., Bradford, Y. M., Conlin, T., et al. (2013) *Nucleic Acids Res.*, **41**, D854–60,
355 ZFIN, the Zebrafish Model Organism Database: increased support for mutants and
356 transgenics.
- 357 2. Bult, C. J., Eppig, J. T., Blake, J. A., et al. (2016) *Nucleic Acids Res.*, **44**, D840–7, Mouse
358 genome database 2016.
- 359 3. Vrandečić, D. and Krötzsch, M. (2014) *Commun. ACM*, **57**, 78–85, Wikidata: a free
360 collaborative knowledgebase.
- 361 4. Quillitz, B. and Leser, U. In *The Semantic Web: Research and Applications*; Bechhofer, S.;
362 Hauswirth, M.; Hoffmann, J.; Koubarakis, M., Eds.; Lecture Notes in Computer Science;
363 Springer Berlin Heidelberg, 2008; pp. 524–538.
- 364 5. Putman, T. E., Burgstaller-Muehlbacher, S., Waagmeester, A., et al. (2016) *Database*,
365 **2016**, Centralizing content and distributing labor: a community model for curating the very
366 long tail of microbial genomes.
- 367 6. Quest, D. J., Land, M. L., Brettin, T. S., et al. (2010) *BMC Bioinformatics*, **11 Suppl 6**, S15,
368 Next generation models for storage and representation of microbial biological annotation.
- 369 7. Miles, A., Zhao, J., Klyne, G., et al. (2010) *J. Biomed. Inform.*, **43**, 752–761, OpenFlyData:
370 an exemplar data web integrating gene expression data on the fruit fly *Drosophila*
371 *melanogaster*.
- 372 8. Cheung, K.-H., Yip, K. Y., Smith, A., et al. (2005) *Bioinformatics*, **21 Suppl 1**, i85–96,
373 YeastHub: a semantic web use case for integrating data in the life sciences domain.
- 374 9. Prud'Hommeaux, E., Seaborne, A. and Others (2008) *W3C recommendation*, **15**, SPARQL
375 query language for RDF.
- 376 10. Pérez, J., Arenas, M. and Gutierrez, C. In *The Semantic Web - ISWC 2006*; Cruz, I.;
377 Decker, S.; Allemang, D.; Preist, C.; Schwabe, D.; Mika, P.; Uschold, M.; Aroyo, L. M.,
378 Eds.; Lecture Notes in Computer Science; Springer Berlin Heidelberg, 2006; pp. 30–43.
- 379 11. Himmelstein, D. Integrating resources with disparate licensing into an open network
380 [https://thinklab.com/discussion/integrating-resources-with-disparate-licensing-into-an-open-](https://thinklab.com/discussion/integrating-resources-with-disparate-licensing-into-an-open-network/107#1)
381 [network/107#1](https://thinklab.com/discussion/integrating-resources-with-disparate-licensing-into-an-open-network/107#1).
- 382 12. Burgstaller-Muehlbacher, S., Waagmeester, A., Mitraka, E., et al. (2016) *Database*, **2016**,
383 Wikidata as a semantic framework for the Gene Wiki initiative.
- 384 13. Elvira Mitraka, Andra Waagmeester, Sebastian Burgstaller, Lynn M. Schriml, Benjamin M.
385 Good, Andrew I. Su *Proceedings of the 2015 Swat4LS International Conference in*
386 *Cambridge England*, Wikidata: A platform for data integration and dissemination for the life
387 sciences and beyond.
- 388 14. Skinner, M. E., Uzilov, A. V., Stein, L. D., et al. (2009) *Genome Res.*, **19**, 1630–1638,
389 JBrowse: a next-generation genome browser.
- 390 15. Buels, R., Yao, E., Diesh, C. M., et al. (2016) *Genome Biol.*, **17**, 66, JBrowse: a dynamic
391 web platform for genome visualization and analysis.
- 392 16. O'Leary, N. A., Wright, M. W., Brister, J. R., et al. (2016) *Nucleic Acids Res.*, **44**, D733–45,
393 Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and
394 functional annotation.
- 395 17. Albrecht, M., Sharma, C. M., Dittrich, M. T., et al. (2011) *Genome Biol.*, **12**, R98, The
396 transcriptional landscape of *Chlamydia pneumoniae*.
- 397 18. Mao, F., Dam, P., Chou, J., et al. (2009) *Nucleic Acids Res.*, **37**, D459–63, DOOR: a
398 database for prokaryotic operons.
- 399 19. Okuda, S. and Yoshizawa, A. C. (2011) *Nucleic Acids Res.*, **39**, D552–5, ODB: a database
400 for operon organizations, 2011 update.
- 401 20. Mons, B., Ashburner, M., Chichester, C., et al. (2008) *Genome Biol.*, **9**, R89, Calling on a
402 million minds for community annotation in WikiProteins.

- 403 21. [Howe, D., Costanzo, M., Fey, P., et al. \(2008\) *Nature*, **455**, 47–50, Big data: The future of](#)
 404 [biocuration.](#)
 405 22. [Pico, A. R., Kelder, T., van Iersel, M. P., et al. \(2008\) *PLoS Biol.*, **6**, e184, WikiPathways:](#)
 406 [pathway editing for the people.](#)
 407 23. [Huss, J. W., 3rd, Orozco, C., Goodale, J., et al. \(2008\) *PLoS Biol.*, **6**, e175, A gene wiki for](#)
 408 [community annotation of gene function.](#)
 409 24. [Hoffmann, R. \(2008\) *Nat. Genet.*, **40**, 1047–1051, A wiki for the life sciences where](#)
 410 [authorship matters.](#)
 411 25. [Lee, E., Helt, G. A., Reese, J. T., et al. \(2013\) *Genome Biol.*, **14**, R93, Web Apollo: a web-](#)
 412 [based genomic annotation editing platform.](#)

413 Figures

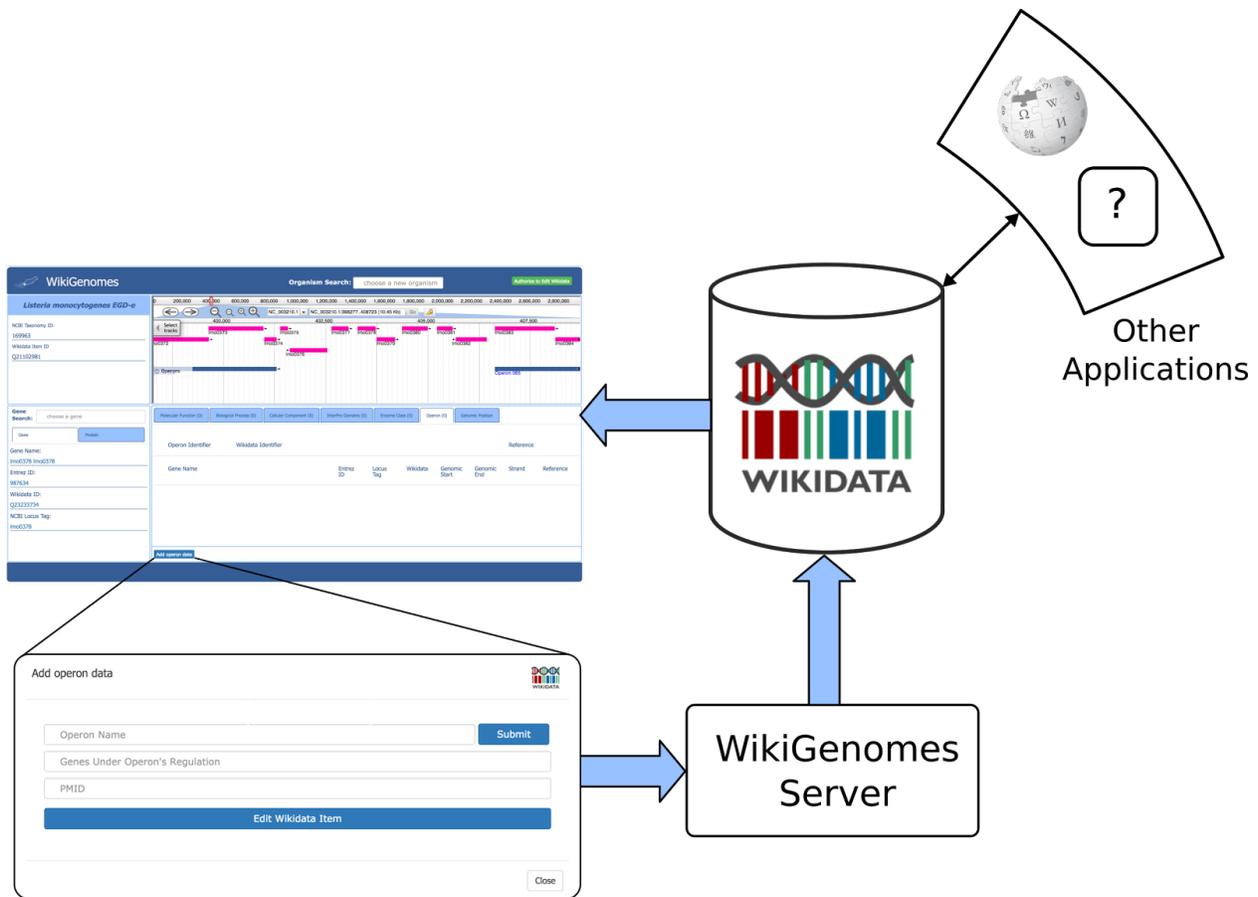
414 **Figure 1**



415 **Figure 1 Wikidata Data Model and Sources** Schematic of the basic structure of the data
 416 model in Wikidata. Entities include purple: Organism data sourced from NCBI's taxonomy
 417 database (www.ncbi.nlm.nih.gov/taxonomy), light pink: gene data sourced from NCBI's Gene
 418 Database (www.ncbi.nlm.nih.gov/gene), green: Gene product data sourced from UniProt
 419 (www.uniprot.org), red: Gene Ontology annotations sourced from EBI's QuickGO API
 420 (www.ebi.ac.uk/QuickGo), orange: InterPro domain annotations sourced from the InterPro
 421 project (www.ebi.ac.uk/interpro) and yellow: Operon data currently sourced from primary
 422 publications hosted in PubMed (www.ncbi.nlm.nih.gov/pubmed). The names of the Wikidata
 423 properties that construct the data model are included in the upper left hand corner.
 424
 425

426 **Figure2**

427



428

429

430 **Figure 2 Process of contributing to and consuming data in WikGenomes via Wikidata.**

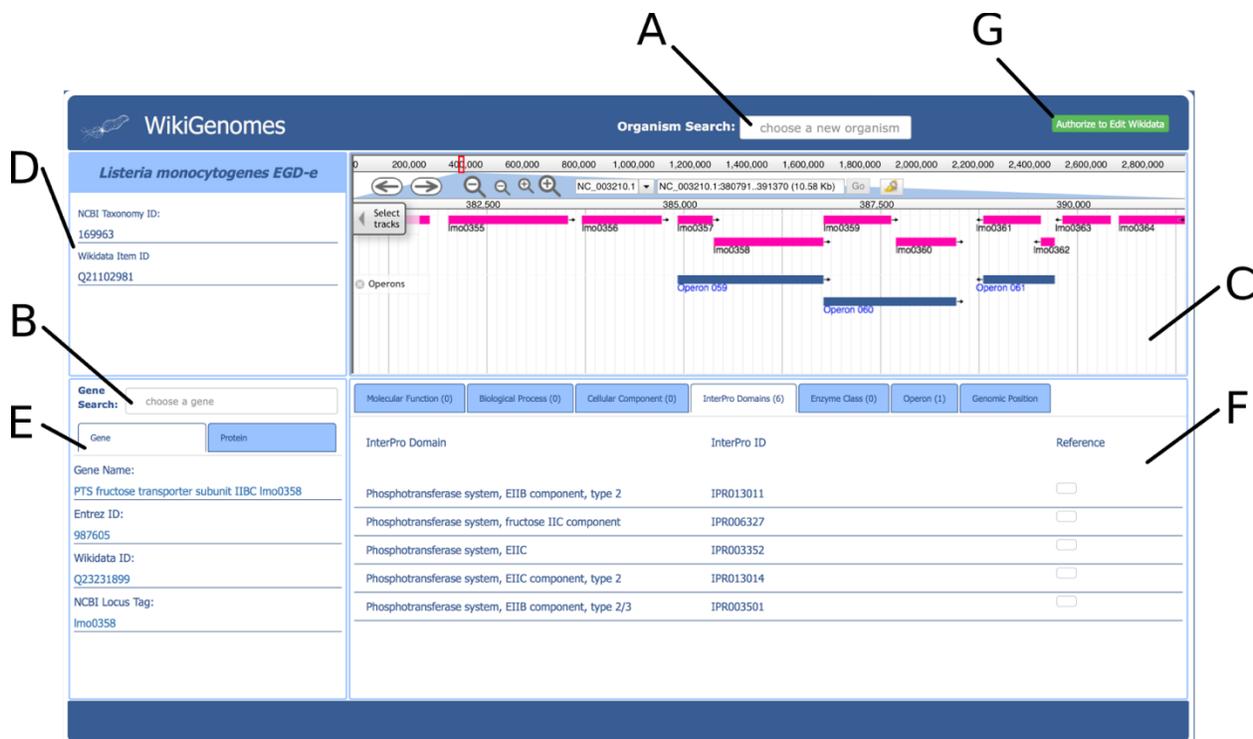
431 WikiGenomes retrieves data from Wikidata via the WDQS SPARQL query service, and allows

432 contribution of defined edits back into Wikidata via a guided annotation curation process. The

433 data then becomes available to any web application that uses Wikidata in a similar way.

434

435 Figure 3



436

437

438 **Figure 3 Overview of the WikiGenomes Interface.** A. The ‘Organism Search’ form selects

439 the organism and populates the main page of the application with data specific to that

440 organism. B. The ‘Gene Search’ form loads data for a selected gene/protein. C. **Genome** The

441 WikiGenomes genome browser is an instance of JBrowse, a high-performance, web-based,

442 client-side genome browser that currently displays gene and operon tracks. D. **Organism** The

443 “Organism Box’ displays the name of the selected organism and basic core identifiers

444 (hyperlinked to their respective database entries). These include the Wikidata ‘QID’, and the

445 taxonomy ID from NCBI’s Taxonomy Database. This content window is where any type of

446 metadata about the organism will be added as it becomes available in Wikidata (e.g.,

447 morphology, lifestyle, Gram staining, disease associations, drugs that have action against it). E.

448 **Gene/Protein** The ‘Gene/Protein’ content window displays the gene name and basic identifiers

449 for the currently loaded gene and gene product in a tabbed view. The ‘Gene’ tab includes the

450 NCBI Entrez ID, the Wikidata QID, and the NCBI Locus Tag (a core identifier for bacterial

451 genes). The ‘Protein’ tab contains the UniProt ID, NCBI RefSeq Protein ID and Wikidata

452 QID. These are the core identifiers required for the current data model in Wikidata, but as more

453 and more mapped identifiers from different resources are added to Wikidata, they will be

454 displayed here. F. **Annotations** The annotations that have been curated in Wikidata for the

455 currently loaded Gene/Protein are displayed in the content window in the bottom right corner of

456 the application, separated and organized by navigation tabs. The tabs represent the domains of

457 annotations that are pulled from various resources into Wikidata for bacterial genes and

458 proteins. Currently, annotation types include Molecular Function, Biological Process, Cellular

459 Component, InterPro Domain, Operon, Enzyme Commission Number and Genomic Position.

460 Each annotation type displays its own relevant information consisting of hyperlinked database

461 identifiers, the hyperlinked Wikidata identifier, the annotation itself, and (if appropriate) genomic

462 coordinates of the annotation (which are also rendered in the genome browser). All annotations

463 are linked to references retrieved from their underlying Wikidata statements. G. **Authorize to**

464 **Edit Button.** This button redirects to WikiMedia.org login where the user can user login to their
465 Wikidata account and authorize WikiGenomes to make edits using their credentials.
466

467 **Figure 4**

468

A

Add a Molecular Function to this protein

Gene Ontology Term

Determination Method ▾

PMID

Edit Wikidata Item

Close

B

Add operon data

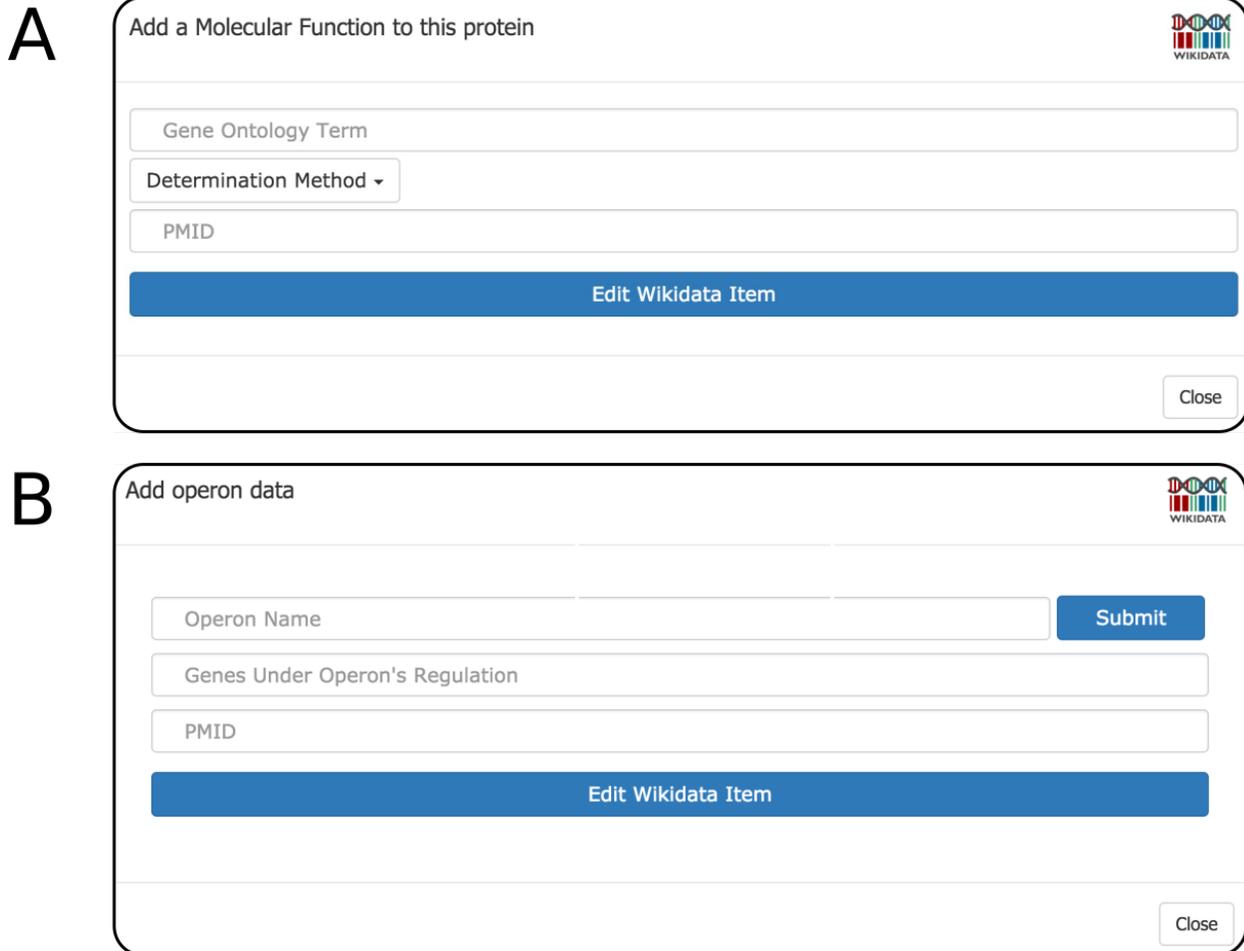
Operon Name

Genes Under Operon's Regulation

PMID

Edit Wikidata Item

Close

Figure 4 consists of two panels, A and B, each showing a web form for editing Wikidata items. Panel A, titled 'Add a Molecular Function to this protein', contains a text input for 'Gene Ontology Term', a dropdown menu for 'Determination Method', and another text input for 'PMID'. Below these is a large blue button labeled 'Edit Wikidata Item' and a 'Close' button in the bottom right corner. Panel B, titled 'Add operon data', features a text input for 'Operon Name' followed by a blue 'Submit' button, a text input for 'Genes Under Operon's Regulation', and a text input for 'PMID'. It also includes a large blue 'Edit Wikidata Item' button and a 'Close' button in the bottom right corner. Both forms include a Wikidata logo in the top right corner.

469

470

471 **Figure 4 Editing Forms A. Gene Ontology Form** The Gene Ontology Form prompts the user
472 to supply 3 pieces of information: 1) the Wikidata item of the GOterm, 2) the method that was
473 used for determination (GO evidence code *i.e. derived from experiment, sequence similarity,*
474 *etc...*) and 3) the PMID of the publication that the statement originated from. The GO term
475 selection box incorporates type-ahead autocomplete allowing the user to find and select the
476 proper GO term to describe the annotation. Upon user submission, WikiGenomes submits the
477 API call to write the new annotation to Wikidata. **B. Operon Form** The Operon Form prompts
478 the user to enter the name of the operon; if the operon already exists, WikiGenomes will find it
479 in Wikidata and add to it. If the operon does not exist in Wikidata, the new name will be set as
480 the label of a new Wikidata item for that operon. The user then provides the genes whose
481 expression is regulated by the operon. The input field for genes also provides type-ahead
482 functionality, allowing the user to quickly search for and select genes from that genome. Like all

483 WikiGenomes forms, the user must also provide the publication that the statement was derived
484 from.
485
486