Published online 5 December 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D9–D12***
doi:10.1093/nar/gkr1195

Making your database available through Wikipedia: the pros and cons

Robert D. Finn[1,*], Paul P. Gardner[2] and Alex Bateman[3]

[1]HHMI Janelia Farm Research Campus, 19700 Helix Drive, Ashburn, VA, USA, [2]School of Biological Sciences, University of Canterbury, Christchurch 8140, New Zealand and [3]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK

Received November 15, 2011; Accepted November 16, 2011

ABSTRACT

Wikipedia, the online encyclopedia, is the most famous wiki in use today. It contains over 3.7 million pages of content; with many pages written on scientific subject matters that include peer-reviewed citations, yet are written in an accessible manner and generally reflect the consensus opinion of the community. In this, the 19th Annual Database Issue of *Nucleic Acids Research,* there are 11 articles that describe the use of a wiki in relation to a biological database. In this commentary, we discuss how biological databases can be integrated with Wikipedia, thereby utilising the pre-existing infrastructure, tools and above all, large community of authors (or Wikipedians). The limitations to the content that can be included in Wikipedia are highlighted, with examples drawn from articles found in this issue and other wiki-based resources, indicating why other wiki solutions are necessary. We discuss the merits of using open wikis, like Wikipedia, versus other models, with particular reference to potential vandalism. Finally, we raise the question about the future role of dedicated database biocurators in context of the thousands of crowdsourced, community annotations that are now being stored in wikis.

INTRODUCTION

Over the recent years of the NAR Database Issue, there has been an increasing trend of articles being submitted where the database makes substantial use of a wiki. A wiki is a web-based tool for the curation and editing of a set of web pages. So, why have wikis become so popular? This trend is not specific to biological databases and the popularity of wikis stems from their simplicity and availability. A wiki provides a simple framework for capturing and sharing of data, generated by any user with a web browser and the appropriate permissions to edit the wiki content. Indeed, Ward Cunningham, developer of the first wiki software, described it as being 'the simplest online database that could possibly work' (see http://en.wikipedia.org/w/index.php?title=Wiki&oldid=457195237). The most famous wiki on the web today is Wikipedia, the online encyclopedia. Wikipedia currently contains over 3.7 million pages of content, most of which can be edited by anyone. A large proportion of search engine queries return a Wikipedia article as one of the top hits. Consequently, most users of the web have come across a wiki in some shape or form, even if they did not realise it.

So, how do wikis relate to biological databases? Traditionally, these databases have employed small teams of expert curators, typically based at one (or maybe a few sites), to write the scientific content regarding the entries in the database. However, with the advent of wikis, the curation burden can be distributed to many more people, unrestricted by geographical location, or simply 'crowdsourced'. Moreover, wikis allows the scientists who are experts on a given topic to be engaged and to share their knowledge, rather than a biological curator having to generate content based on published literature. The openness of wikis seems odd to many scientists who are used to the peer-review model used for publications. Some scientists may assume, incorrectly, that because anyone can edit a wiki at any time, the content must be flawed, especially if they do not know who that user is.

WIKIS AND BILOGICAL DATABASES

Vandalism and the danger of open wikis

For Pfam and Rfam we adopted the use of Wikipedia as the wiki model. One of the main reasons that we came to this decision is because we believed that the openness and profile of Wikipedia embraces the highest number of potential editors. As there is no need to log-in, users can more readily edit the text, even for simple typographical corrections. There is then instant user gratification of seeing the change appear in the article. Furthermore,

*To whom correspondence should be addressed. Tel: +01223 495330; Fax: +01223 494919; Email: finnr@janelia.hhmi.org

this approach takes advantage of 'hot topics' as they suddenly spike in public interest. Huss *et al*. (1) elegantly showed the correlated trend between hot topics in the news (as measured by Google searches) and the number of edits in Wikipedia in relation to the GeneWiki project. We have observed similar trends between the release of prominent papers and the number of edits to pages pertaining to those articles.

However, the open nature is also a potential drawback of using Wikipedia, as anyone can freely edit the content, in any way they see fit. Articles in Wikipedia have little provenance on each editor, and the lack of editing restrictions increases the likelihood that content will be inappropriately edited. Inappropriate editing falls into two broad categories, (i) blatant, malicious defacing of the content or addition of spam links and advertising; and (ii) the more subtle errors introduced, either intentionally or due to mis-understanding, resulting in misleading content. The former are obvious to someone reading the content and may cause offence. This sort of editing is a common problem for high profile wikis, such as Wikipedia. However, due to the large community of editors and the use of bots (programs that automatically edit pages) most vandalism of this nature is promptly reverted. It is the second type that is more worrying to a scientist. For example, on 3 June 2007, an anonymous Wikipedian (a user who has not logged in) modified the article on the regulatory RNA 'Riboswitches' to include a reference to an episode of the cartoon 'The Simpsons' that featured Ribwiches. It was not until 17 June 2007 that this was removed. It is unclear whether this was done maliciously, or whether it was caused by a confusion between Ribwiches and Riboswitches. Overall, inappropriate editing is rare for pages related to biology and occurs at very low rates. Since Rfam started to use Wikipedia to manage the textual annotation of RNA families in 2007, ~1% of all edits have been reverted, by the Rfam curators or the greater Wikipedia community, suggesting that they may be vandalism (2). Similar numbers are presented for the GeneWiki project (3), with these authors estimating that vandalism is observed less than one page view in every 3000, a rate much lower than other articles in Wikipedia.

For many scientists, a closed wiki (restricted by user login, not to be confused with private, where viewing the content is restricted) may seem to be the obvious choice: who would trust a paper where the authors were anonymous? You want to know the provenance of the edit and to know that the user has been vetted in some way to ensure they have the appropriate scientific background. While this undoubtedly helps ensure the validity of the content, there is an 'energy barrier' to overcome. Any new user must first obtain a username and password and this initial, albeit small, energy barrier will deter many casual browsers who might only want to make a small change. This undoubtedly restricts the user base, probably to those who have a vested interest in the database—this may be the desired effect. With a defined user group, it is significantly easier to enforce more consistency in the layout, media and textual content of each wiki page.

An intermediate solution between open and closed wikis has been adopted by some of the wiki databases that are published in this issue. In this model, although the wiki is closed there is a relatively low barrier to being accepted as an editor. In the 'vampire' model adopted by GONUTS (4) and EcoliWiki (5) any registered user can create new user accounts for their colleagues. Thus, the burden of maintaining the editor list is shared among all editors.

There are 11 different articles in this Database Issue of NAR that describe the use of wikis as all or part of the database, listed in Table 1. Despite the use of different wiki implementations and the varied data that are being annotated, those articles reporting numbers of wiki edits are typically reporting thousands of edits per year (3–5,6). These numbers are consistent with those that we have observed (2).

Interestingly, both resources that utilise Wikipedia describe independently developed quality control procedures for Wikipedia articles. Pfam (and Rfam) uses the Wikipedia API to track new edits and present them to the biocurators for approval to ensure that the changes to the article are appropriate, before the article is displayed on the database website (7). The authors of the GeneWiki project have developed the WikiTrust resource (3), which works via a Firefox plug-in, to mark up Wikipedia articles according to the Wikipedian's reputation. Both approaches have their merits and are both aimed at maintaining quality in the open wiki setting.

Database integration levels with Wikipeida

There are many ways that a database can be integrated with Wikipedia. The GeneWiki project is completely contained within Wikipedia as a portal, which acts like a homepage to the project—there is no standalone database. Each gene annotated by the project corresponds to a full article. But, full articles are not the only way to contribute to Wikipedia and expose biological databases. For example, the 'infoboxes' found on the right side on many Wikipedia pages provide a more structured way of adding data. The PDB database of protein structures (8) has a large number of links from infoboxes in Wikipedia. So, although it would not make sense to have a Wikipedia article about any particular protein crystal structure, the information is very relevant to readers of articles about biological molecules. Biological databases such as UniProtKB (9) are increasingly adding links to Wikipedia articles because the information is recognized as valuable. Pfam and Rfam adopt a model where having identified relevant articles in Wikipedia, we not only link to them, but we also import the articles back into our web pages. This is made possible by the open license used by Wikipedia. If a user wishes to change the annotation that Pfam or Rfam displays then they only need to go to Wikipedia to edit that article.

Leveraging the Wikipedia community

There are a number of excellent forums for discussing how to incorporate your ideas into Wikipedia. This is highly recommended before embarking on a mass of changes that may not be appropriate for Wikipedia but may be

Table 1. A list of the databases utilising wikis that appear in this issue

Resource	URL	Brief description
EcoliWiki	http://ecoliwiki.net	*Escherichia coli* strain, phage, plasmid, and mobile genetic element information resource
GeneWiki	http://en.wikipedia.org/wiki/Portal:Gene_Wiki	Collection of human gene annotations
GONUTS	http://gowiki.tamu.edu	Gene ontology application guide
Metadatabase	http://metadatabase.org	Catalogue of biological databases
959 Nematode genomes	http://www.nematodes.org	Nematode sequencing project resource
Pfam	http://pfam.janelia.org; http://pfam.sanger.ac.uk	Database of protein families, with annotations stored in Wikipedia
SEQwiki	http://SEQwiki.org	Catalogue of tools, technologies and tutorials for high-throughput sequencing
SNPedia	http://www.SNPedia.com	Resource of the functional consequences of human genetic variation
SubtiWiki	http://subtiwiki.uni-goettingen.de	Collection of *Bacillus subtilis* genome annotations
WikiPathways	http://www.wikipathways.org	Pathway curation and annotations
XanthusBase	http://www.xanthusbase.org	Annotation resource for *Myococcus xanthus* and related bacteria

tweaked and modified in order to make a positive contribution. Many well-meaning editors are put off by negative responses and even wholesale removal of their initial edits and articles. For example, if you added hundreds of links to your database from Wikipedia this is likely to be viewed simply as advertising. By engaging the existing communities these problems can be avoided, vastly improving the editing experience. Of particular importance to the readers here are the WikiProjects: 'Molecular and Cellular Biology' and 'Computational Biology' and the 'Gene Wiki' Portal (3). Wiki-colleagues there will be happy to make useful suggestions and may get involved in your project. Keep in mind that your main focus should be improving Wikipedia and not promoting your sphere of influence or getting more people to view your product. It is worth starting by editing some articles to get an idea of how Wikipedia works, as well as building up a history of trusted edits. Wikipedia editors will take you more seriously if you are seen as one of them. For more advice on getting involved in Wikipedia there is a 10 simple rules article with plenty of useful advice (10).

Wikipedia infrastructure and content

In this section, we are not planning on comparing different wiki applications, which should be based on the requirements of the database, for example, database storage engine, programming language, file uploads and so on (see http://en.wikipedia.org/wiki/Comparison_of_wiki_software). Rather, can an existing wiki, such as Wikipedia be used as the wiki, rather than duplicating software and hardware and going through the hardship of maintaining the infrastructure yourself?

This decision is almost certainly going to be influenced by whether an open or closed user base is required. If the database wants to control the user access, then that database will have to maintain the wiki. However, other factors will influence whether Wikipedia is appropriate. Not all content is appropriate for Wikipedia, as facts need to be cited to a reliable, published source (http://en.wikipedia.org/wiki/Wikipedia:Verifiability) and the content needs to be devoid of jargon and accessible to a broad audience. For example, the wiki TOPSAN (11) includes user's conjecture as to the function of certain protein domains, based on unpublished structures and other experiments, this sort of content would violate Wikipedia's policy regarding 'No Original Research' and would be removed (http://en.wikipedia.org/wiki/Wikipedia:No_original_research).

Other wikis, such as PDBwiki (12) and SNPedia (13) contain data that are too specific to be included in an encyclopedia such as Wikipedia. Therefore, both the type of content and granularity of entries need to be carefully considered before using Wikipedia. Concepts that are unlikely to be considered notable by the broader Wikipedia community should not have their own entries. For example, each of the 24 534 SNPs in SNPedia or the 75 245 structures in PDBwiki are not significant in their own right. However, these data sets may make excellent sections on less specific articles, such as on a notable gene with SNPs influencing phenotype and structures elucidating function. Furthermore, SNPedia does not maintain Wikipedia's 'No original research' policy and instead actively invites well-documented original research. Two important questions that you need to ask when thinking about adding to Wikipedia are: are your data notable and verifiable? And, should it be merged with existing Wikipedia entries rather than creating new and overly specific entries?

The limitations of wikis in relation to biological databases

Are wikis the future of biological databases? Well, the answer is yes, and no. They help bridge the vast gap between the amount of human-annotated data compared to unannotated data. Where annotations do exist, the use of wikis can help to improve the annotation and keep it relevant. But, wikis are extremely simple and are unlikely to replace the complex relational organisation of many biological databases. Furthermore, when using Wikipedia, the database that uses a particular article has no control over future revisions even though they may have created it. We have already mentioned editing, but there is a more substantial control issue. An article, if the Wikipedia community considers it appropriate, can be deleted or merged into another article. Although

page-redirects are maintained, the article may be substantially different to the original article that was cited. In our experience, the deleting or merging of entries is a rare event and tends to improve content for articles used by Pfam or Rfam. However, this goes back to the point of thinking about granularity and content of the article in Wikipedia in the first place.

In the case of Pfam and Rfam, there is a logical marriage with Wikipedia that stores the annotations for the families. All the other family data, such as multiple sequence alignments are stored in a local MySQL database. As well as the relational data model, wikis also do not lend themselves to storing data entities such as alignments, structures and probabilistic models. Most databases have extensive quality control procedures, which do not fit the simple edit and save model of the wiki. Consequently, other mechanisms are required to add new entries to the database [for an example from Pfam, see ref. (7)], but wikis, in whatever form, provide an excellent platform for storing free text annotations that are editable by a distributed user community.

Is the work of a dedicated database biocurator at an end?

There is a risk to the funding of biological databases and biocurators that can be created by the implementation of successful community annotation. Will funders and grant reviewers think that biological databases can be curated by a diffuse network of volunteers? This is certainly not the case and at the core of every successful wiki database are a group of dedicated experts who do the bulk of the data curation. One of the lessons from wiki databases is that there is a large community of people who are prepared to make a small number of changes to their subject of interest. But this piecemeal contribution will not give a comprehensive set of data that is required for biological integration of data. However, we envisage that the role of biocurators may change. Instead of solely curating entries, they will need to train new users, verify edits and resolve conflicts as they arise.

CONCLUSION

Wikis are undoubtedly changing the way some biological databases operate, providing an established solution to community annotation. Adopting a wiki means that a particular resource is no longer a closed, static resource (between public data releases) that cannot be improved by its users. The challenge is now to get scientists *en masse* to generate and edit articles. How editors receive credit for their work on the article is unclear. Assuming they work on the subject area, wiki articles provide them an opportunity to showcase their work in context of the field at the very least. Curation of biological data must be crowdsourced if there is any chance to comprehensively annotate the vast datasets that are being generated and the scientific community should feel responsible. Hopefully this article has provoked thoughts as to how a wiki, especially Wikipedia, may work for a resource that you are responsible for or use. The growing number of databases using wikis suggests that they are here to stay, we now face the issue of how to overcome the social engineering required to get everyone involved.

FUNDING

Howard Hughes Medical Institute (to R.D.F); Rutherford Discovery Fellowship from NZ Government (to P.P.G.); Wellcome Trust (WT098051 to A.B.). The open access publication charge for this paper has been waived by Oxford University Press—*NAR* Editorial Board members are entitled to one free paper per year in recognition of their work on behalf of the journal.

Conflict of interest statement. None declared.

REFERENCES

1. Huss,J.W., Lindenbaum,P., Martone,M., Roberts,D., Pizarro,A., Valafar,F., Hogenesch,J.B. and Su,A.I. (2010) The Gene Wiki: community intelligence applied to human gene annotation. *Nucleic Acids Res.*, **38**, D633–D639.
2. Gardner,P.P., Daub,J., Tate,J., Moore,B.L., Osuch,I.H., Griffiths-Jones,S., Finn,R.D., Nawrocki,E.P., Kolbe,D.L., Eddy,S.R. *et al.* (2011) Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res.*, **39**, D141–D145.
3. Good,B.M., Clarke,E.L., de Alfaro,L. and Su,A.I. (2011) The Gene Wiki in 2011: community intelligence applied to human gene annotation. *Nucleic Acids Res.*, **40**, D1255–D1261.
4. Renfro,D., McIntosh,B., Venkatraman,A., Siegele,D. and Hu,J.C. (2011) GONUTS: The Gene Ontology Normal Usage Tracking System. *Nucleic Acids Res.*, **40**, D1262–D1269.
5. McIntosh,B., Renfro,D., Knapp,G., Lairikyengbam,C., Liles,N., Niu,L., Supak,A., Venkatraman,A., Zweifel,A., Siegele,D. *et al.* EcoliWiki: A Wiki-based community resource for *Escherichia coli*. *Nucleic Acids Res.*, **40**, D1270–D1277.
6. Li,J.-W., Robison,K., Martin,M., Sjödin,A., Usadel,B., Young,M.D., Olivares,E. and Bolser,D. (2011) The SEQanswers wiki: a wiki database of tools for high throughput sequencing analysis. *Nucleic Acids Res.*, **40**, D1313–D1317.
7. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2011) The Pfam Protein Families database. *Nucleic Acids Res.*, **40**, D290–D301.
8. Rose,P.W., Beran,B., Bi,C., Bluhm,W.F., Dimitropoulos,D., Goodsell,D.S., Prlić,A., Quesada,M., Quinn,G.B., Westbrook,J.D. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.
9. UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
10. Logan,D.W., Sandal,M., Gardner,P.P., Manske,M. and Bateman,A. (2010) Ten simple rules for editing Wikipedia. *PLoS Comput. Biol.*, **6**, e1000941.
11. Ellrott,K., Zmasek,C.M., Weekes,D., Sri Krishna,S., Bakolitsa,C., Godzik,A. and Wooley,J. (2011) TOPSAN: a dynamic web database for structural genomics. *Nucleic Acids Res.*, **39**, D494–D496.
12. Stehr,H., Duarte,J.M., Lappe,M., Bhak,J. and Bolser,D.M. (2010) PDBWiki: added value through community annotation of the Protein Data Bank. *Database*, doi: 10.1093/database/baq009.
13. Lennon,G. and Cariaso,M. SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Res.*, **40**, D1308–D1312.