PERSPECTIVE

nature genetics

A wiki for the life sciences where authorship matters

Robert Hoffmann

WikiGenes is the first wiki system to combine the collaborative and largely altruistic possibilities of wikis with explicit authorship. In view of the extraordinary success of Wikipedia there remains no doubt about the potential of collaborative publishing, yet its adoption in science has been limited. Here I discuss a dynamic collaborative knowledge base for the life sciences that provides authors with due credit and that can evolve via continual revision and traditional peer review into a rigorous scientific tool.

In WikiGenes (http://www.wikigenes.org/), authorship tracking technology enables users to directly identify the source of every word. This was not possible in first generation wikis, although authorship is essential to acknowledge contributors and to appraise the reliability of information. On the basis of clear authorship attribution, users can rate each other, and a self-regulating reputation system can be implemented. This is useful to address quality maintenance and the problem of editing conflicts, which used to depend on slow and theoretically refutable top-down decisions. To facilitate contribution and unambiguous use of scientific language, WikiGenes enables editing of articles in their final layout and citation of scientific terminology and references through integrated database and ontology lookups. All contributions to WikiGenes will be open access.

Dynamic publishing

Wikipedia is global, easy to use, and has a low barrier to access. Many web-based applications have this in common, however. What is outstanding about the wiki model is of course the collaboration of many authors and, moreover, the way in which collaboration manifests itself in the content. A comparison with conventional publications makes this clear.

The medium of conventional publications has changed over thousands of years of human history, but the essential characteristics have not. Conventional publications have a limited number of authors, a precise date of publication, and thus a definite version. In one word, conventional publications are static[1]. In science, this means that scholarly discourse must take place in a series of static publications, spread over different journals and media.

Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, Massachusetts 02139, USA. Correspondence should be addressed to R.H. (hoffmann@wikigenes.org).

Published online 27 August 2008; doi:10.1038/ng.f.217

Dynamic publications in the wiki model, on the other hand, have theoretically an unlimited number of authors, but more importantly, they have no final version and no definite date of publication. Dynamic articles can evolve with the content and their focus may shift over time. In the context of science, dynamic publications could thus integrate scientific discourse and be potentially always up to date. In a scientific wiki, for instance, there would be no explicit errata, only improved versions of an article.

Continuous integration and harmonization of scientific discourse may not always be desirable and would probably impede scientific innovation and progress. But often it is necessary to synthesize novel insights and theories and to create common reference points. Reviews and textbooks fulfill this function in the conventional system.

The comparison of static and dynamic publishing models shows that they are not mutually exclusive but complementary in the best sense. Dynamic publications, however, are a recent phenomenon, and we are currently exploring whether science can benefit from their integrative potential. This is why so many wikis have been created over the past years in and out of science. By now there are wikis for different scientific communities and on a growing number of topics, ranging from quantum information science to neuropsychiatry. The acceptance of many of these novel approaches in the scientific community has been low, however, and skepticism persists[2]. One reason may simply be that the integrative forces of a bottom-up approach are not useful in all contexts. Most important, first generation wikis (**Box 1**) have not been created specifically for the demands in science, and significant technical innovation is required to unleash the potential of dynamic publishing for scientific discourse.

WikiGenes

WikiGenes is a collaborative knowledge resource for the life sciences, which is based on the general wiki idea but employs specifically developed technology to serve as a rigorous scientific tool. The rationale behind WikiGenes is to provide a platform for the scientific community to collect, communicate and evaluate knowledge about genes, chemicals, diseases and other biomedical concepts in a bottom-up process.

The necessity of this approach originates in the fast-growing body of information in biology and medicine. Over the past decades, a hypercycle of technological advancement, scientific hypothesis and increasing amount of data has led to a shift in perspective toward 'system thinking'[3,4]. Genome-wide experiments and pathway screenings, for instance, confront researchers with genes or chemical substances they might never have heard of[5]. To sustain scientific progress, it is thus important to provide researchers access to overview information on all biological agents and systems. For this reason, large institutions, such as the National Center for Biotechnology Information (NCBI) and the

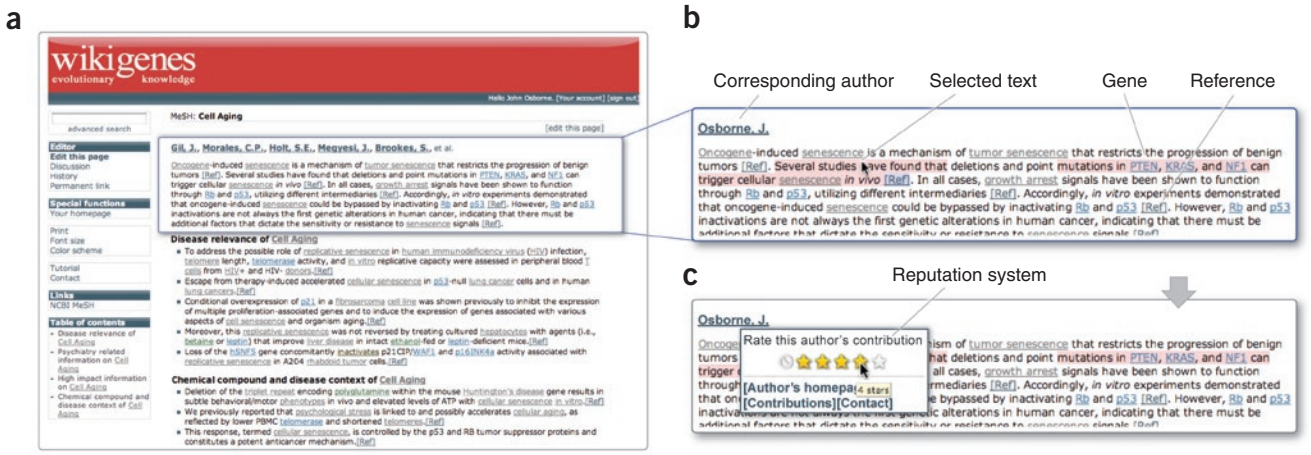

Figure 1 Authorship tracking and reputation system. (**a**) Front page of a typical article in WikiGenes (http://www.wikigenes.org/). Every word in the article is linked unambiguously to its contributing author. (**b**) Clicking in the text of interest highlights the corresponding author and all of her or his contributions (pink). This way, authors, who invest time and knowledge in their contributions, are given due credit. Readers, on the other hand, can always know the author of any part of a collaborative article in WikiGenes. Hyperlinks on scientific terminology (genes, for example) and references are provided throughout the system. (**c**) The author name is linked to a context menu, providing access to the author's contributions in other articles and additional information useful to appraise the author. On the basis of this clear authorship attribution, users can rate each other, and it is possible to implement a self-regulating reputation system. This way, readers may acknowledge contributions of particular authors, and dedicated authors could enhance their reputation and assume more responsibility, for instance, in the settlement of editing conflicts.

European Bioinformatics Institute (EBI), invest extensively in expert-curated information resources[6–9]. These top-down approaches are, however, extremely time consuming and expensive and can ultimately not keep pace—nor scale—with the rapid increase of information[10,11]. Use of a collaborative community-driven approach, on the other hand, would make it possible to assemble a scientific encyclopedia for biomedicine that would integrate all aspects, be always up to date and evolve with our collective level of knowledge.

Even though there is no doubt as to the usefulness of such a global, community-driven resource in principle[12–14], it should not be ignored that the expected effort from contributing scientists would be considerable. This is especially true in a competitive field such as biomedicine, where academic and professional success depends significantly on publication records[15]. To make scientific wikis viable, it is therefore essential to exploit all technical possibilities to minimize the necessary efforts for contributors and, more importantly, to provide authors due recognition for their contributions.

Authorship matters

Curiosity and the satisfaction of working together and exchanging ideas are outstanding human characteristics. Another essential element to being a scientist is recognition by others, which translates into employment, grants and, ultimately, the privilege of being a scientist[16]. Recognition in science is closely linked to authorship. How many scientists would publish their research and best theories in *Nature* without authorship, anonymously?

In addition to prestige, authorship provides a basis to establish priority of discoveries and theories and to build a reputation among peers. The lack of clear authorship in first-generation wikis (**Box 1**) is therefore detrimental to the usefulness of dynamic publishing in science. Moreover, in the context of scientific wikis, it is only fair to duly acknowledge authors, who invest time and knowledge in their contributions.

In WikiGenes, authorship tracking technology is used to link every contribution unambiguously to its author, creating the first hybrid of traditional, scientific and collaborative, dynamic publishing (**Fig. 1**). This technical innovation in WikiGenes also supports the other central function of authorship as guidance for the reader. Authorship is essential to appraise origin, authority and reliability of information. This is especially important in the wiki model, with its dynamic content and large number of authors. The quality of Wikipedia, for instance, is often so good[17] that it is easy to forget how collaborative articles come about. An author creates a paragraph, another deletes a sentence and inserts a word here and there, and a third author moves a paragraph and adds a new aspect. In brief, the history of a collaborative article can become extremely complex within a few editing cycles.

How could the reader of such an article know who wrote what? In first generation wikis, this information can theoretically be found in the archives and attempts have been made to establish reliability measures[18], but in practice, it is impossible for a user to reconstruct the authorship of specific text passages from hundreds of previous versions.

The uncertainty as to the source of specific texts is therefore an important problem in dynamic publications and decreases the value of articles in their entirety[19]. In WikiGenes, on the contrary, new contributions are identified with every editing step and attributed to their authors. Thus readers can always know the corresponding author of any part of a WikiGenes article.

Reputation system and community-driven review

Having a large number of authors is vital to the integration of diverse viewpoints and the efficient assembly of an extensive body of knowledge. Yet, it also harbors the risk of varying quality, tendentious argumentation and, in the worst case, vandalism[17,19]. In conventional publishing, editorial peer review warrants the adherence to standards, a preselection of information, and a certain level of quality[20,21]. Scientific wikis should be able to benefit from this time-tested formula.

It is in fact possible to adopt essential elements of editorial peer review in the wiki model. The large number and changeability of dynamic publications and the effort expected from editors and reviewers, however, make conventional peer review difficult to scale. Therefore, it is important that the entire community of authors and readers exerts review and quality control[22]. In wikis, every author may also act as reviewer, critically questioning or improving disputed information or adding crucial

references. In this approach, reviewers and authors are working together to improve the quality and reliability of articles.

Clear authorship in WikiGenes facilitates this review process by providing information on the source and, thus, on the likely relevance and reliability of specific texts. One click in the text highlights the corresponding author, linked to all of her contributions and other information useful to appraise the author. Hence, the overall value of a collaborative article is less affected by the varying quality of contributions.

Linking texts unambiguously to their authors also makes it possible for authors in WikiGenes to rate each other on the basis of specific contributions. Thus, for the first time, collaborative publishing can be enhanced with the advantages of a fine-grained reputation system[23–26] (**Fig. 1**). This way, readers may acknowledge contributions of particular authors, and all users can benefit from each other's experience[24]. Furthermore, the reputation system in WikiGenes enables readers to report author-specific vandalism and therefore also functions as an immune system, preserving the interests of the community. On the other hand, especially committed authors could enhance their reputation and assume more responsibility[27,28], for instance, in the arbitration of editing conflicts[17,19].

The initial reputation system in WikiGenes is straightforward, but it has the potential to evolve with increasing user experience[25]. To catalyze the reputation system, for instance, it would be possible to transfer reputation from the real to the virtual world, for example, by considering the conventional publication success of individual authors, or via user registration with a recognized organization, such as HUGO[29,30]. It is important to stress, however, that the implementation of a reputation system in the wiki model would not be conceivable without unambiguous authorship.

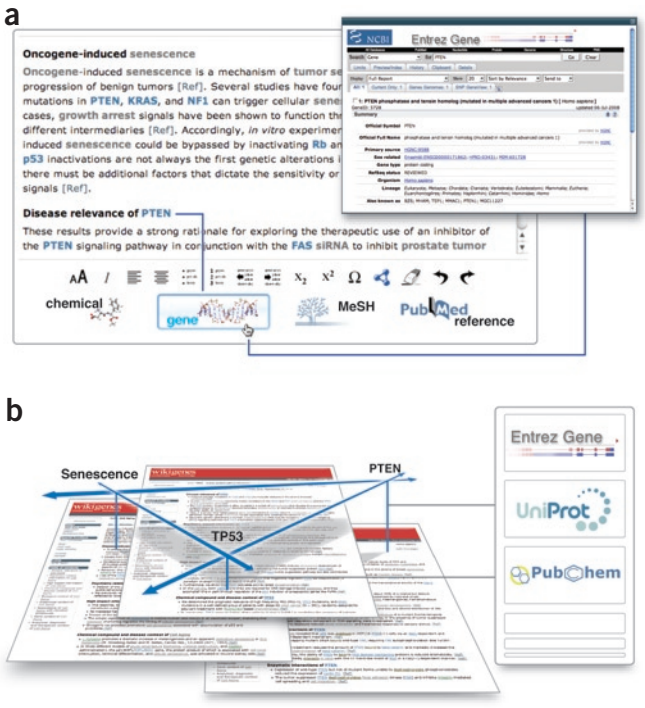

Figure 2 Citation of scientific terminology and references. (**a**) WikiGenes provides a specialized editor for easy editing of articles and the management of scientific terminology and references. The editor facilitates the creation and modification of articles in their final layout, avoiding a complicated syntax. Unambiguous citation of scientific terminology is facilitated through integrated database and ontology lookups. It is possible, for instance, to link the term 'PTEN' to the specific human gene without the need to know the exact database number. (**b**) Users can benefit from this meta-information by using genes and biomedical concepts as hyperlinks between articles. This way, the content in WikiGenes becomes accessible as one navigable network of information and is also connected to primary resources such as the NCBI or UniProt.

Box 1 Evolution of wiki technology

First generation wikis. The first wiki software, WikiWikiWeb, was written in 1994 by Ward Cunningham. Since then, dozens of Wiki software implementations have been created, for instance MediaWiki (http://www.mediawiki.org/), which is used by Wikipedia. A number of biologically relevant wikis are based on this technology; for example, SNPedia (http://www.snpedia.com/) uses the same technology as Wikipedia for core wiki functionalities. In Wikiproteins (a collective approach to link information in a machine-readable resource), natural language wikis are secondary and also based on MediaWiki.

Wikis with authorship and reputation system. Unambiguous authorship information is essential to acknowledge contributors and to appraise the reliability of information. In first-generation wikis, authorship can theoretically be found in the archives, but in practice, it is impossible for a user to reconstruct it from hundreds of previous versions. WikiGenes is the first wiki to provide unambiguous authorship attribution and thus a basis for reputation systems.

Citation of scientific terminology and references

The electronic representation of articles on the Internet makes it possible to dynamically organize and link information. This is especially useful in the life sciences, where it is increasingly problematic to describe complex processes within the linear constraints of written text. Our understanding of signal transduction, for instance, has evolved from linear cascades to dynamic complex formation and pathway crosstalk[31–33], which are difficult to represent in non-electronic publications without hyperlinks. Hyperlinks are increasingly needed to refer to unique entities or precise insertion points within larger documents, meaning that content-specific interlinking is wreaking irreversible change to publishing practices, perhaps to the point where it will soon no longer make sense to print an article, as essential hyperlink functionality and metadata would get lost[34,35].

Wikis, on the other hand, are at home on the Internet, and intuitive linking of articles is one of their strengths. In common wiki systems, any word can easily be made into a hyperlink to an article with this word as a title. Yet in science, this kind of linking is too unspecific because scientific language is dense with specialized terminology. Particularly in biology and medicine, thousands of ambiguous synonyms and abbreviations are used for diseases, genes and chemicals. The term 'p53', for instance, may refer to genes from different organisms.

To fully exploit the potential of hyperlinks, it is therefore important to unambiguously define scientific terms using database and ontology identifiers[36]. In most wiki systems, this is only possible by using an arbitrary editing syntax and the knowledge of relevant database numbers. Given the enormous size of biological databases, this would be a bothersome hurdle for many potential users. For this reason, WikiGenes provides a specialized editor for easy editing of articles and the management of scientific terminology and references (**Fig. 2**). Avoiding the use of a complicated syntax, the editor facilitates the creation and modification of articles in their final layout. Besides standard editing functionality, integrated database and ontology lookups are supported to unambiguously identify and cite scientific concepts and publications. This way, it

is possible, for instance, to link the term 'p53' to the specific human *TP53* gene without the need to know the exact database number.

This meta-information is useful to link to external resources, such as the NCBI or UniProt, and to integrate textual content with experimental data[6,8,37–40]. Moreover, other users can benefit from this information by using genes and biomedical concepts as hyperlinks between articles[36]. This way, the entire content in WikiGenes becomes accessible as an interlinked and navigable information resource, capable to represent complex biological processes.

Community project

Open collaborations, such as the Human Genome Project, would not have been possible without the Internet, distributed systems for the aggregation, review and dissemination of knowledge and, most importantly, the active support of a large community[41]. One of the next ambitious goals is the analysis of genetic variation in all human genes and its impact on specific diseases and pathologies[42,43]. The Human Variome Project, for instance, is an international initiative to gather and curate data on all human genetic variation in a globally accessible form[43]. A demanding project of this dimension cannot succeed without a collective effort of the entire community of researchers and clinicians.

A comprehensive collection of mutations and phenotypes is, apart from the scientific challenge, also hindered by the fact that it is increasingly difficult to publish mutation data in a conventional way[43]. For the progress of this field, it is thus essential to find novel incentives and possibilities[34] to ensure that no result remains unpublished. In this context a collaborative publishing platform with clear authorship attribution can play an essential role. It enables the assembly of every bit of information into an overall picture, and it also provides novel ways for scientists to closely interact, make contacts and coordinate efforts[2,44].

To support collaborative assembly of knowledge in an early stage, WikiGenes provides more than a hundred thousand automatically generated articles on chemical compounds, proteins, organisms, pathologies and other biomedical concepts[45]. These articles were compiled from the iHOP information resource (http://www.ihop-net.org/) and are organized in subsections to cover functional and regulatory relationships, disease relevance and more[36,46]. These germinal articles do not come near the quality of expert-curated information, but they fulfill an important function as a substrate and matrix to embed contributions in the early collaborative process.

WikiGenes was endorsed by the Human Variome Project and will support this community effort from the beginning. Yet, the scope of WikiGenes is—despite its name—much broader. In fact, all researchers, clinicians, teachers and other professionals are invited to contribute to a common scientific knowledge base of biology and medicine. All contributions to WikiGenes will be open access in perpetuity.

Future prospects

The technological innovation in WikiGenes is central to the attempt to turn the wiki model into a rigorous scientific tool. To this aim it is also important to provide a framework that supports the contribution of novel and original research. Clear authorship attribution facilitates this essentially, but the integrative and harmonizing forces in dynamic publications tend to work against original and novel views. In WikiGenes, authors are therefore provided with the option to create protected articles with a limited number of selected co-authors. These articles cannot be edited by others, but they can still be linked to the encyclopedic core and discussed and rated by everyone. This way, it would be possible in the near future to publish original research and establish priority of discoveries and theories.

Besides the ubiquitous community-driven review, another important objective in WikiGenes is to facilitate formal peer-review by established experts. Formal peer-review is essential to prevent systematic bias, detect weaknesses or outright errors and establish a reliable body of research and knowledge. In the context of scientific wikis, the implementation of a two-tier review model seems natural. Articles with high access rates or community ratings could be selected by editors and submitted for in-depth review. Community efforts, such as the Human Variome Project, or health advocacy groups[47] could gain from the comprehensiveness of collaborative approaches and support specific reviews. Established journals, on the other hand, could harvest hot topics from scientific wikis and commission peer-reviewed overviews[48]. Thus, a close integration of dynamic and conventional publishing would be possible, getting the best of both worlds: creative collaboration and static, high-quality reference articles.

This is perhaps the place to emphasize that the possibilities for scientific collaboration shown in WikiGenes are not limited to biology. Other fields, such as physics, can benefit from collaborative publishing with clear authorship, reputation system, and integrated ontologies. Edit this article.

ACKNOWLEDGMENTS

This work was supported in part by the Branco Weiss Fellowship, Society in Science. I am thankful to the Memorial Sloan-Kettering Cancer Center for hosting the iHOP information resource, C. Sander and M. Ramirez-Gaite for helpful discussion, and T. Berners-Lee and D. Weitzner for their support.

Published online at http://www.nature.com/naturegenetics/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Butler, D. Science in the web age: joint efforts. *Nature* **438**, 548–549 (2005).
2. Waldrop, M.M. Science 2.0. *Sci. Am.* **298**, 68–73 (2008).
3. Westerhoff, H.V. & Palsson, B.O. The evolution of molecular biology into systems biology. *Nat. Biotechnol.* **22**, 1249–1252 (2004).
4. Taipale, J. & Beachy, P.A. The Hedgehog and Wnt signalling pathways in cancer. *Nature* **411**, 349–354 (2001).
5. Patterson, S.D. & Aebersold, R.H. Proteomics: the first decade and beyond. *Nat. Genet.* **33** (Suppl.), 311–323 (2003).
6. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35** (Database issue), D61–D65 (2007).
7. Brooksbank, C., Cameron, G. & Thornton, J. The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Res.* **33** (Database issue), D46–D53 (2005).
8. The UniProt Consortium. The universal protein resource (UniProt). *Nucleic Acids Res.* **36** (Database issue), D190–D195 (2008).
9. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33** (Database issue), D514–D517 (2005).
10. Baumgartner, W.A. Jr, Cohen, K.B., Fox, L.M., Acquaah-Mensah, G. & Hunter, L. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics* **23**, i41–i48 (2007).
11. Salzberg, S.L. Genome re-annotation: a wiki solution? *Genome Biol.* **8**, 102 (2007).
12. Wang, K. Gene-function wiki would let biologists pool worldwide resources. *Nature* **439**, 534 (2006).
13. Mons, B. *et al.* Calling on a million minds for community annotation in WikiProteins. *Genome Biol.* **9**, R89 (2008).
14. Huss, J.W. *et al.* A gene wiki for community annotation of gene function. *PLoS Biol.* **6**, e175 (2008).
15. Check, E. More biologists but tenure stays static. *Nature* **448**, 848–849 (2007).
16. Lehmann, S., Jackson, A.D. & Lautrup, B.E. Measures for measures. *Nature* **444**, 1003–1004 (2006).
17. Giles, J. Internet encyclopaedias go head to head. *Nature* **438**, 900–901 (2005).
18. Adler, B.T., de Alfaro, L. A content-driven reputation system for the Wikipedia in *Proc. WWW.*, 261–270 (ACM Press, New York, 2007).
19. Giles, J. Wikipedia rival calls in the experts. *Nature* **443**, 493 (2006).
20. Spier, R. The history of the peer-review process. *Trends Biotechnol.* **20**, 357–358 (2002).
21. Kassirer, J.P. & Campion, E.W. Peer review. Crude and understudied, but indispensable. *J. Am. Med. Assoc.* **272**, 96–97 (1994).
22. Cokol, M., Iossifov, I., Rodriguez-Esteban, R. & Rzhetsky, A. How many scientific papers should be retracted? *EMBO Rep.* **8**, 422–423 (2007).
23. Conte, R. & Paolucci, M. *Reputation in Artificial Societies: Social Beliefs for Social Order.* (Kluwer Academic Publishers, Boston, 2002).

24. Resnick, P., Zeckhauser, R., Friedman, E. & Kuwabara, K. Reputation systems. *Comm. ACM Conf.* **43**, 45–48 (2000).
25. Friedman, E. & Resnick, P. The social cost of cheap pseudonyms. *J. Econ. Manage. Strategy* **10**, 173–199 (2001).
26. Kling, R. *Computerization and Controversy: Value Conflicts and Social Choices* 2nd edn. (Academic Press, San Diego, 1996).
27. Chen, M. & Singh, J.P., Computing and using reputations for internet ratings, in *Proc. ACM Conf. Electronic Commerce*. 154–162 (ACM Press, New York, 2001).
28. Rodriguez, M.A., Bollen, J. & Van de Sompel, H. The convergence of digital libraries and the peer-review process. *J. Inf. Sci.* **32**, 149–159 (2006).
29. McKusick, V.A. HUGO news. The Human Genome Organisation: history, purposes, and membership. *Genomics* **5**, 385–387 (1989).
30. Little, P. Human genome annotation—a possible role for HUGO? Human Genome Organisation. *Nat. Genet.* **19**, 222 (1998).
31. de Lichtenberg, U., Jensen, L.J., Brunak, S. & Bork, P. Dynamic complex formation during the yeast cell cycle. *Science* **307**, 724–727 (2005).
32. Campbell, S.L., Khosravi-Far, R., Rossman, K.L., Clark, G.J. & Der, C.J. Increasing complexity of Ras signaling. *Oncogene*, **17** (11 Reviews),1395–1413 (1998).
33. von Bubnoff, A. & Cho, K.W. Intracellular BMP signaling regulation in vertebrates: pathway or network? *Dev. Biol.* **239**, 1–14 (2001).
34. Anonymous. Compete, collaborate, compel. *Nat. Genet.* **39**, 931 (2007).
35. Heber, J. Print and perish? *Nat. Mater.* **7**, 512–514 (2008).
36. Hoffmann, R. & Valencia, A. A gene network for navigating the literature. *Nat. Genet.* **36**, 664 (2004).
37. Berners-Lee, T. & Hendler, J. Publishing on the semantic web. *Nature* **410**, 1023–1024 (2001).
38. Searls, D.B. Data integration: challenges for drug discovery. *Nat. Rev. Drug Discov.* **4**, 45–58 (2005).
39. Anonymous. Let data speak to data. *Nature* **438**, 531 (2005).
40. Wheeler, D.L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **36** (Database issue), D13–D21 (2008).
41. Roberts, L., Davenport, R.J., Pennisi, E. & Marshall, E. A history of the Human Genome Project. *Science* **291**, 1195 (2001).
42. Frazer, K.A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
43. Cotton, R.G. *et al.* Recommendations of the 2006 Human Variome Project meeting. *Nat. Genet.* **39**, 433–436 (2007).
44. Anonymous. Join a social revolution. *Nature* **436**, 1066 (2005).
45. Lipscomb, C.E. Medical Subject Headings (MeSH). *Bull. Med. Libr. Assoc.* **88**, 265–266 (2000).
46. Hoffmann, R. & Valencia, A. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics* **21** (Suppl. 2), ii252–ii258 (2005).
47. Terry, S.F., Terry, P.F., Rauen, K.A., Uitto, J. & Bercovitch, L.G. Advocacy groups as research organizations: the PXE International example. *Nat. Rev. Genet.* **8**, 157–164 (2007).
48. Anonymous. What is the human variome project? *Nat. Genet.* **39**, 423 (2007).