PLOS | BIOLOGY

PERSPECTIVE

# Biocuration: Distilling data into knowledge

**International Society for Biocuration**[¶]*

¶ Contributions from members of the International Society for Biocuration are provided in the Acknowledgments.
* intsocbio@gmail.com

## Abstract

Data, including information generated from them by processing and analysis, are an asset with measurable value. The assets that biological research funding produces are the data generated, the information derived from these data, and, ultimately, the discoveries and knowledge these lead to. From the time when Henry Oldenburg published the first scientific journal in 1665 (*Proceedings of the Royal Society*) to the founding of the United States National Library of Medicine in 1879 to the present, there has been a sustained drive to improve how researchers can record and discover what is known. Researchers' experimental work builds upon years and (collectively) billions of dollars' worth of earlier work. Today, researchers are generating data at ever-faster rates because of advances in instrumentation and technology, coupled with decreases in production costs. Unfortunately, the ability of researchers to manage and disseminate their results has not kept pace, so their work cannot achieve its maximal impact. Strides have recently been made, but more awareness is needed of the essential role that biological data resources, including biocuration, play in maintaining and linking this ever-growing flood of data and information. The aim of this paper is to describe the nature of data as an asset, the role biocurators play in increasing its value, and consistent, practical means to measure effectiveness that can guide planning and justify costs in biological research information resources' development and management.

OPEN ACCESS

**Citation:** International Society for Biocuration (2018) Biocuration: Distilling data into knowledge. PLoS Biol 16(4): e2002846. https://doi.org/10.1371/journal.pbio.2002846

**Published:** April 16, 2018

**Funding:** The author(s) received no specific funding for this work.

**Competing interests:** The authors have declared that no competing interests exist.

**Abbreviations:** DOI, digital object identifier; EMBO, European Molecular Biology Organization; ENCODE, Encyclopedia of DNA Elements; FAIR, findable, accessible, interoperable, and reproducible; FAIR-TLC, findable, accessible, interoperable, reproducible, traceable, appropriately licensed, and interconnected; GA4GH, Global Alliance for Genomics and Health; GEO, Gene Expression Omnibus; IMEx, International Molecular Exchange; ISB, International Society for Biocuration; NCBI, National Center for Biotechnology Information; ORCID, Open Researcher and Contributor ID; PDF, portable

## Data as an asset

Research data continue to be produced at ever-growing rates due to both technological advances and decreasing costs for their generation [1]. Understanding what makes data assets distinct from other types of assets is fundamental in terms of their valuation and effective management [2]. To briefly summarise, from an economic perspective, its unique characteristics are these: Information is infinitely shareable without any decrease in its intrinsic value. For example, the same sequence retrieved from the National Center for Biotechnology Information (NCBI) can be shared by an unlimited number of people without any loss of value. Unlike physical assets—e.g., sequencing equipment, which depreciates with use—information sharing actually increases its value in a compound fashion; and reciprocally, unshared information is

document format; UMLS, Unified Medical Language System; wwPDB, Worldwide Protein Database.

**Provenance:** Not commissioned; externally peer reviewed.

less valuable [3,4,5]. Further, the more accurate and complete the information is, the more valuable it is. In other words, quality is at least as important as quantity [6,7,8]. Since inferences are only as good as the information they are based upon, inaccuracies and omissions compel scientists to spend valuable research time winnowing out poor-quality or inaccurate information or, even worse, inadvertently ploughing research funds into dead ends. Moreover, with the increasing role of automatic inference systems for high-throughput data and data analytics, there is a growing dependency on the availability of robust, high-quality knowledge resources, and the gold-standard data sets they contain, for benchmarking. Lastly, when information is combined, its value increases. For example, genetic testing can reveal hundreds of thousands of variants per individual, yet for most variants, the clinical consequences are not yet known [9]. If our goal is to advance research, instantiation of known connections is essential to accelerate the process of relating genotypes to phenotypes in a way that is impossible when using individual data sets in isolation [10,11,12,13,14].

Managing a biological information resource relies on a range of intersecting skills: Bioinformaticians, application developers, system administrators, biocurators, journal editors, etc. are all involved in this collective effort. Within this context, biocurators focus on information content rather than technology. Their overarching goal is to maximise the value of the information assets researchers are generating by assuring their accuracy, comprehensiveness, integration, accessibility, and reuse.

## What is biocuration?

Biocuration is the extraction of knowledge from unstructured biological data into a structured, computable form. In this context, knowledge is most commonly extracted from published manuscripts, as well as from other sources such as experimental data sets and unpublished results from data analysis. Biocurators are typically PhD-level biologists, often with lab bench experience coupled with specialised expertise in computational knowledge representation. Their work entails the synthesis and integration of information from multiple sources—including, for example, peer-reviewed papers; large-scale projects, such as the Encyclopedia of DNA Elements (ENCODE); or conference abstracts. They contact authors directly for clarification, digest supplemental information, and resolve identifiers in order to accurately capture a researcher's conclusion and their evidence for that conclusion. Biocurators strive to distil the current 'best view' from conflicting sources and ensure that their resources provide data that are not only findable, accessible, interoperable, and reproducible (FAIR), but also traceable, appropriately licensed, and interconnected (collectively, the FAIR-TLC principles [15]).

## Biocuration motivation

Scientific communication is shifting in this 'information age', with researchers increasingly relying on curated resources [16,17,18,19]. For example, when comparing an entry in the Worldwide Protein Data Bank (wwPDB; https://www.wwpdb.org)—a resource containing detailed reviewed information on macromolecular structures—with a portable document format (PDF) file containing a figure of the same structure, it is obvious that the latter, non-computer-readable representation is insufficient for downstream comparative use. The political processes in the scientific community that led to designating wwPDB [20], the International Nucleotide Sequence Database Collaboration [21], and others such as the International Molecular Exchange (IMEx) [22] and ProteomeXchange consortia [23] as official depositories have proven to be well worth the effort. These examples highlight the importance of collaboration and synergy between journal editors and databases. The definition of what it means to publish is expanding [24], since results only published as a PDF have limited accessibility. To promote impact and reuse, the full semantic spectrum must be employed, from human-readable language to fully computationally interpretable.

## Biocuration costs

Although expert biocuration is clearly labour intensive, it scales surprisingly well with the growth of biomedical literature, as demonstrated by two recent studies [25,26]. Advanced tools are also increasing efficiency and accuracy, and biocurators are often actively engaged as team members in developing machine learning and natural-language processing techniques. Although these methods currently lack the necessary precision and recall required for a real-world setting [27,28,29], they are beginning to provide assistance [30,31,32,33,34] and will continue to incrementally improve.

The costs for sustaining a useful research resource in which biocuration plays an essential role represent only a tiny fraction of the original research funding [35]. An independent survey assessing the value of biological database services concluded that the benefits to users and their funders are equivalent to more than 20 times the direct operational cost of the institute [36]. Additionally, the hidden cost of an individual researcher's time spent trawling the literature to find the information pertinent to their own specialist field is impossible to estimate, but having the required data easily accessible in a structured format represents a considerable saving in person-hours and, therefore, money for every funder, academic institute, and biomedical enterprise.

## Actionable recommendations

### Everyone can be a biocurator—Data reporting fit for knowledge synthesis

Seriously addressing seemingly mundane issues—such as identifying gene symbols, isoforms, strains, antibodies, and cell lines—is essential if experimental results are to be correctly integrated within the existing body of knowledge. For example, a recent study found that almost 40% of the gene lists submitted to the Gene Expression Omnibus (GEO) and 20% of the gene lists in the supplementary material of published articles contain gene symbol errors introduced by the software used during data handling prior to publication [37]. This will continue to be a significant problem until infrastructure is in place at key junctions in the research life cycle. New tools and workflows are needed for connecting researchers, journals, reviewers, and repositories and easily conveying standards-compliant information.

Progress is being made; notably, community guides for provisioning and referencing life science identifiers have recently been published [38,39], outlining best practices for facilitating large-scale data integration. Likewise, in the lab, software applications that support autocompletion within individual cells of spreadsheets, as well as more sophisticated standards-aware data collection tools, ensure that standard terminologies are applied as data are collected [40,41,42]. Through the use of such electronic laboratory notebook and manuscript submission software and the adoption of recommended formats and community-endorsed terminologies and ontologies, the goal of 'born computable' lab data generation will be realised. Initiatives have also started in scientific journals. A good example is provided by SourceData, a project initiated by the European Molecular Biology Organization (EMBO) press, which involves the biocuration of article figures prior to publication [43].

## Support for standards—Development, usage, and sustainability

Common standards for describing and classifying biology are indispensable for reproducible interactions, information exchange, interoperability, comparability, and discoverability [44]. Without standards, database search results will inevitably miss key information or include irrelevant material.

Biocurators regularly lead efforts in standards development: engaging with experts, building consensus, fostering adoption, and maintaining biological fidelity. Yet apart from a very limited number of cases, funding for standards development is unavailable. Even in the case of the Gene Ontology Consortium [45], the funding for this indispensable standard is significantly aided through other projects. On the other side of the spectrum, the Human Phenotype Ontology [46,47,48] operates using donated time from a handful of dedicated individuals, despite its widespread adoption (e.g., the Unified Medical Language System [UMLS], United Kingdom 100,000 Genomes Project, and the Global Alliance for Genomics and Health [GA4GH]). While the lack of dedicated funding poses a risk, the harmful consequences of not using any standard are vastly greater.

More can be done to inform and educate data producers and consumers on the importance of standards to ensure research data are not wasted or lost in the wrong format, with the wrong metadata descriptions, or described using a private or personal set of terms. Efforts such as FAIRsharing [30] (fairsharing.org), which maps the landscape of databases and standards and links them to the journal and funder data policies that endorse their use, go a long way to making sure that existing standards are adopted. However, more funding is needed for these infrastructure projects to aid data and knowledge sharing, to minimise the duplication of effort, and to ensure that researchers can easily employ appropriate standards.

## Expediting the collection and processing of data

Recently, there has been considerable excitement about the strategy of crowdsourcing, putting biocuration tools into a researcher's hands so that they may directly contribute and publish their results into knowledge resources [49,50,51,52]. There is a tremendous potential in this approach, but to ensure success, there are clear prerequisites that must be satisfied—(i) editorial oversight, (ii) automated integrity checks, and (iii) citation mechanisms. Successful community-sourced projects universally include editorial control, which is where biocurators can play a key role, to avoid collecting poor-quality data that would decrease the value of a resource overall.

In addition, support for developing user interfaces, batch submission tools, and utilities to computationally validate content—such as simple checks for syntactical correctness, falling outside standard deviations, or using disallowed values—is needed for direct data submission. Here again, biocurators often play a role in defining validation standards. Machine-readable standards are critical in this step, as they enable validation to be carried out programmatically. Continuous integration and contextual analysis approaches may even suggest what a contributor might do to improve their data before making a final submission. Notably, biologists are already beginning to use community curation tools when they are available, such as Canto [53]—which is used by researchers working on *Schizosaccharomyces pombe* to directly submit their data to a resource—and Apollo [54], which is used for community-based curation of gene structures for improving automated gene sets.

Lastly, citation mechanisms need to be built into the contribution process. This both acts as an incentive and fosters reproducibility, since information is traceable to the original experimental work that led to a conclusion. Currently, existing biological data resources associate every assertion they contain with its underlying experimental justification by linking it to a

PubMed identifier, which is an indirect route to the actual researcher(s) who contributed this information. Literature citations are mere proxies for assessing productivity and impact. Embedding a traceable authorship facility directly into laboratory software or a resource's submission software would provide a much more direct and accurate means of assessing a researcher's impact. By associating a researcher (e.g., an Open Researcher and Contributor ID [ORCID] persistent identifier, https://orcid.org/) with an identified piece of information (e.g., a persistent identifier, such as a digital object identifier [DOI]), their contributions become citable objects [55,56,57], and the subsequent use of this information by other researchers can be tracked. If this is encouraged, one can envision a time when community curation tools become the first place for digitally publishing research conclusions, shared directly into digital community resources.

## Biocuration is a necessity for scientific progress

Actively promoting innovations in fundamental data and information capture will yield enormous return on our research investment. The existing pain points—the time wasted by individual researchers discovering information, collecting it, manually verifying it, and integrating it in a piecemeal fashion—all impede scientific advancement. For researchers, biocuration means they can easily find extensive and interlinked information at well-documented, stable resources. It means they can access this information through multiple channels by browsing websites, downloading it from repositories, or retrieving it dynamically via web services. It likewise means the information will be as accurate and reliable as possible. And—because biocurators have integrated information by describing it using community semantic standards, applying authoritative identifiers, and transforming it into standard formats—disparate data sets collected from multiple research projects can be directly compared.

## Acknowledgments

Most instrumental for the composition of this paper are the members of the International Society for Biocuration (ISB) themselves, whose valuable contributions, helpful comments, sound suggestions, and eye for details helped to shape this article. Special thanks go to these ISB members, in alphabetical order: Mais Ammari, Andrew Chatr Aryamontri, Helen Attrill, Amos Bairoch, Tanya Berardini, Judith Blake, Qingyu Chen, Julio Collado, Delphine Dauga, Joel T. Dudley, Stacia Engel, Ivan Erill, Petra Fey, Richard Gibson, Henning Hermjakob, Gemma Holliday, Doug Howe, Chris Hunter, David Landsman, Ruth Lovering, Deepa Manthravadi, Aron Marchler-Bauer, Beverley Matthews, Ellen M. McDonagh, Birgit Meldal, Gos Micklem, Daniel Mietchen, Christopher J. Mungall, Kim Pruitt, Vidhya Sagar Rajamanickam, James M. Reecy, Alix Rey, Khader Shameer, Aleksandra Shipitsyna, Ana Luisa Toribio, Mary Ann Tuli, Peter Uetz, Ulrike Wittig, Valerie Wood, and all the many other ISB members (biocuration.org).

For their contribution, we recognise previous members of the ISB executive committee: Teresa Attwood, Alex Bateman, Tanya Berardini, Lydie Bougueleret, Pascale Gaudet, Jennifer Harrow, Tadashi Imanishi, Renate Kania, Lorna Richardson, Marc Robinson-Rechavi, Owen White, Ioannis Xenarios, and Chisato Yamasaki.

Appreciation for steering the manuscript forward goes to the following members of current and recent ISB executive committees: Cecilia N. Arighi, Rama Balakrishnan, J. Michael Cherry, Melissa Haendel, Suzanna E. Lewis, Peter McQuilton, Monica Muñoz-Torres, Claire O'Donovan, Sandra Orchard, Sylvain Poux, Andrew Su, Nicole Vasilevsky, and Zhang Zhang.

## References

1. Kahn SD. On the Future of Genomic Data. Science 2011;  331 (6018), 728–9 https://doi.org/10.1126/science.1197891 PMID: 21311016
2. Moody D., Walsh P., Measuring the Value of Information: an Asset Valuation Approach, presented at European conference on Information Systems, June 1999
3. Glazer R. (1993) Measuring the Value of Information: The Information Intensive Organisation. IBM Systems Journal,  Vol 32, No 1, 99–110.
4. Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE  2(3): e308. https://doi.org/10.1371/journal.pone.0000308 PMID: 17375194
5. Anagnostou P, Capocasa M, Milia N, Sanna E, Battaggia C, Luzi D, et al. (2015) When Data Sharing Gets Close to 100%: What Human Paleogenetics Can Teach the Open Science Movement. PLoS ONE 10(3): e0121409. https://doi.org/10.1371/journal.pone.0121409 PMID: 25799293
6. Dasu T, Johnson T. Exploratory data mining and data cleaning.  John Wiley & Sons; 2003 Aug 15.
7. Feldman B, Martin EM, Skotnes T. Big Data in Healthcare Hype and Hope. 2012 Oct;360. [cited 2016]. https://www.ghdonline.org/uploads/big-data-in-healthcare_B_Kaplan_2012.pdf
8. Hazen BT, Boone CA, Ezell JD, Jones-Farmer LA. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. International Journal of Production Economics. 2014 Aug 31; 154:72–80
9. Landrum M.J., Lee J.M., Riley G.R., Jang W., Rubinstein W.S., Church D.M., et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014; 42:D980–D985 https://doi.org/10.1093/nar/gkt1113 PMID: 24234437
10. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, et al. Clinical assessment incorporating a personal genome. Lancet. 2010 May 1; 375(9725):1525–35. https://doi.org/10.1016/S0140-6736(10)60452-7 PMID: 20435227
11. Li L, Cheng WY, Glicksberg BS, Gottesman O, Tamler R, Chen R, Bottinger EP, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Sci Transl Med. 2015 Oct 28; 7(311):311ra174. https://doi.org/10.1126/scitranslmed.aaa9364 PMID: 26511511
12. Bone WP, Washington NL, Buske OJ, Adams DR, Davis J, Draper D, et al. Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. Genet Med. 2016 Jun; 18(6):608–17. https://doi.org/10.1038/gim.2015.137 PMID: 26562225.
13. McMurry JA, Köhler S, Washington NL, Balhoff JP, Borromeo C, Brush M, et al. Navigating the Phenotype Frontier: The Monarch Initiative. Genetics. 2016 Aug; 203(4):1491–5. https://doi.org/10.1534/genetics.116.188870 PMID: 27516611.
14. Shameer K, Tripathi LP, Kalari KR, Dudley JT, Sowdhamini R. Interpreting functional effects of coding variants: challenges in proteome-scale prediction, annotation and assessment. Brief Bioinform. 2016 Sep; 17(5):841–62. https://doi.org/10.1093/bib/bbv084 PMID: 26494363
15. Haendel M, Su A, McMurry J, Chute CG, Mungall C, Good B, et al. Metrics to assess value of biomedical digital repositories: response to RFI NOT-OD-16-133. Zenodo; Geneva: 2016
16. Bourne P. Will a Biological Database Be Different from a Biological Journal? PLoS Comput Biol. 2008 1(3): e34. https://doi.org/10.1371/journal.pcbi.0010034
17. Salimi N, Vita R. The biocurator: connecting and enhancing scientific data. PLoS Comput Biol. 2006 Oct 27; 2(10):e125. Review. https://doi.org/10.1371/journal.pcbi.0020125 PMID: 17069454.
18. Hirschman J, Berardini TZ, Drabkin HJ, Howe D. A MOD(ern) perspective on literature curation. Mol Genet Genomics. 2010 May; 283(5):415–25. https://doi.org/10.1007/s00438-010-0525-8 Epub 2010 Mar 11. PMID: 20221640
19. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, et al. Big data: The future of biocuration. Nature. 2008 Sep 4; 455(7209):47–50. https://doi.org/10.1038/455047a PMID: 18769432.
20. Young JY, Westbrook JD, Feng Z, Sala R, Peisach E, Oldfield TJ. et al. OneDep: Unified wwPDB System for Deposition, Biocuration, and Validation of Macromolecular Structures in the PDB Archive. Structure (London, England: 1993) 2017; 25(3):536–545
21. Cochrane G., Karsch-Mizrachi I., Takagi T. International Nucleotide Sequence Database Collaboration The International nucleotide sequence database collaboration. Nucleic Acids Res. 2016; 44(D1):D48–D50 https://doi.org/10.1093/nar/gkv1323 PMID: 26657633
22. Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S, et al. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. Nat Methods. 2012; 9(4):345–50. https://doi.org/10.1038/nmeth.1931 PMID: 22453911

23. Deutsch EW, Csordas A, Sun Z, Jarnuczak A, Perez-Riverol Y, Ternent T. et al. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition Nucleic Acids Res. 2017; 45(Database issue): D1100–D1106.

24. Beyond the PDF. Nat Methods. 2013 Feb; 10(2):91. PMID: 23479796.

25. Poux S, Arighi CN, Magrane M, Bateman A, Wei CH, Lu Z et al. On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. Bioinformatics (Oxford, England). 2017; 33(21):3454–3460. https://doi.org/10.1093/bioinformatics/btx439 PMID: 29036270

26. Oliver SG, Lock A, Harris MA, Nurse P, Wood V. Model organism databases: essential resources that need the support of both funders and users. BMC Biol. 2016; 14: 49 https://doi.org/10.1186/s12915-016-0276-z PMID: 27334346

27. Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci U S A. 2004 Apr 6; 101 (Suppl 1):5228–35. https://doi.org/10.1073/pnas.0307752101 PMID: 14872004.

28. Hersh W. Evaluation of biomedical text-mining systems: lessons learned from information retrieval. Brief Bioinform. 2005 Dec; 6(4):344–56. PMID: 16420733.

29. Huang CC, Lu Z. Community challenges in biomedical text mining over 10 years: success, failure and the future. Brief Bioinform. 2016 Jan; 17(1):132–44. https://doi.org/10.1093/bib/bbv024 Review. PMID: 25935162.

30. Hirschman L, Burns GA, Krallinger M, Arighi C, Cohen KB, Valencia A, et al. Text mining for the biocuration workflow. Database (Oxford). 2012 Apr 18; 2012:bas020. https://doi.org/10.1093/database/bas020 PMID: 22513129.

31. Arighi CN, Carterette B, Cohen KB, Krallinger M, Wilbur WJ, Fey P, et al. An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. Database (Oxford). 2013 Jan 17; 2013:bas056. https://doi.org/10.1093/database/bas056 PMID: 23327936.

32. Cejuela JM, McQuilton P, Ponting L, Marygold SJ, Stefancsik R, Millburn GH, et al. tagtog: interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles. Database, 1 January 2014, Volume 2014, bau03.

33. Karamanis N, Seal R, Lewis I, McQuilton P, Vlachos A, Gasperin C, et al. Natural Language Processing in aid of FlyBase curators. BMC Bioinformatics, 2008 9:193, https://doi.org/10.1186/1471-2105-9-193 PMID: 18410678

34. Wang Q, S Abdul S, Almeida L, Ananiadou S, Balderas-Martínez YI, Batista-Navarro R, et al. Overview of the interactive task in BioCreative V. Database (Oxford). 2016 Sep 1; 2016. https://doi.org/10.1093/database/baw119 PMID: 27589961.

35. ten Hoopen P, Amid C, Buttigieg PL, Pafilis E, Bravakos P, Cerdeño-Tárraga AM, et al. Value, but high costs in post-deposition data curation. Database (Oxford). 2016; 2016.

36. Beagrie N, Houghton J. The Value and Impact of the European Bioinformatics Institute [Internet]. 2016. [cited 2016]. http://www.ebi.ac.uk/about/news/press-releases/value-and-impact-of-the-european-bioinformatics-institute

37. Ziemann M, Eren Y, El-Osta A. Gene name errors are widespread in the scientific literature. Genome Biol. 2016 Aug 23; 17(1):177. https://doi.org/10.1186/s13059-016-1044-7 PMID: 27552985.

38. Haendel M, Chute C, editors. NIH BD2K Workshop on Community-based Data and Metadata Standards Development: Best practices to support healthy development and maximize impact; 2015 Feb 25–26; Bethesda, MD. NIH workshop report; 2016. https://datascience.nih.gov/sites/default/files/bd2k/docs/ExecSumm_CBDMSworkshopFEB2015.pdf

39. McMurry JA, Juty N, Blomberg N, Burdett T, Conlin T, Conte N, et al. Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLoS Biol. 2017 Jun 29; 15(6):e2001414. https://doi.org/10.1371/journal.pbio.2001414 eCollection 2017 Jun. PMID: 28662064.

40. Hankeln W, Buttigieg PL, Fink D, Kottmann R, Yilmaz P, Glockner FO. MetaBar—a tool for consistent contextual data acquisition and standards compliant submission. BMC Bioinformatics. 2010 Jun 30; 11(1):358.

41. Wolstencroft K, Owen S, Horridge M, Krebs O, Mueller W, Snoep JL, et al. RightField: embedding ontology annotation in spreadsheets. Bioinformatics. 2011 Jul 15; 27(14):2021–2 https://doi.org/10.1093/bioinformatics/btr312 PMID: 21622664

42. Strasser C, Kunze J, Abrams S, Cruse P. DataUp: A tool to help researchers describe and share tabular data. F1000Research. 2014 Sep 12; 3:6. https://doi.org/10.12688/f1000research.3-6.v2 PMID: 25653834

43. Liechti R, George N, Götz L, El-Gebali S, Chasapi A, Crespo I, et al. SourceData: a semantic platform for curating and searching figures. Nature Methods. 2017; 14(11), 1021–1022. https://doi.org/10.1038/nmeth.4471 PMID: 29088127

44. Vasilevsky NA, Brush MH, Paddock H, Ponting L, Tripathy SJ, Larocca GM, et al. On the reproducibility of science: unique identification of research resources in the biomedical literature. PeerJ. 2013 Jan; 1: e148. https://doi.org/10.7717/peerj.148 PMID: 24032093

45. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May; 25(1):25–9.

46. Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. 2014 Jan; 42(Database issue):D966–74. https://doi.org/10.1093/nar/gkt1026 Epub 2013 Nov 11. PMID: 24217912.

47. Groza T, Köhler S, Moldenhauer D, Vasilevsky N, Baynam G, Zemojtel T, et al. The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease. Am J Hum Genet. 2015 Jun 24; 97 (1):111–24. https://doi.org/10.1016/j.ajhg.2015.05.020 PMID: 26119816

48. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, et al. The Human Phenotype Ontology in 2017. Nucleic Acids Res. 2017 Jan 4; 45(D1):D865–76. https://doi.org/10.1093/nar/gkw1039 PMID: 27899602

49. Karp PD. Crowd-sourcing and author submission as alternatives to professional curation. Database (Oxford). 2016; 2016. https://doi.org/10.1093/database/baw149 PMID: 28025340

50. Khare R, Good BM, Leaman R, Su AI, Lu Z. Crowdsourcing in biomedicine: challenges and opportunities. Brief Bioinform. 6 Jan; 17(1):23–32. https://doi.org/10.1093/bib/bbv021 Epub 2015 Apr 17. Review. PMID: 25888696.

51. McQuilton P, Gonzalez-Beltran A, Rocca-Serra P, Thurston M, Lister A, Maguire E, et al. BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. Database (Oxford). 2016; 2016. https://doi.org/10.1093/database/baw075 PMID: 27189610

52. Lintott CJ, Schawinski K, Slosar A, Land A, Bamford S, Thomas D, et al. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey, Monthly Notices of the Royal Astronomical Society, Volume 389, Issue 3, 21 September 2008, Pages 1179–1189, https://doi.org/10.1111/j.1365-2966.2008.13689.x

53. Rutherford KM, Harris MA, Lock A, Oliver SG, Wood V. Canto: an online tool for community literature curation. Bioinformatics. 2014 Jun 15; 30(12):1791–2. https://doi.org/10.1093/bioinformatics/btu103 PMID: 24574118

54. Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, et al. Web Apollo: a web-based genomic annotation editing platform. Genome Biol. 2013 Aug 30; 14(8):R93. https://doi.org/10.1186/gb-2013-14-8-r93 PMID: 24000942.

55. Tsueng G, Good BM, Ping P, Golemis E, Hanukoglu I, van Wijnen AJ, et al. Gene Wiki Reviews—Raising the quality and accessibility of information about the human genome. Gene. 2016 Nov 5; 592 (2):235–8. https://doi.org/10.1016/j.gene.2016.04.053 PMID: 27150585

56. Piwowar HA, Vision TJ. Data reuse and the open data citation advantage. PeerJ. 2013 Oct 1; 1:e175. https://doi.org/10.7717/peerj.175 PMID: 24109559

57. Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, et al. Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Comput Sci. 2015 Jan 27; 1:e1. https://doi.org/10.7717/peerj-cs.1 PMID: 26167542