PLOS BIOLOGY

ESSAY

From Reductionism to Reintegration: Solving society's most pressing problems requires building bridges between data types across the life sciences

Anne E. Thessen[1]*, Paul Bogdan[2], David J. Patterson[3], Theresa M. Casey[4], César Hinojo-Hinojo[5], Orlando de Lange[6], Melissa A. Haendel[1]

1 Department of Environmental and Molecular Toxicology, Oregon State University, Corvallis, Oregon, United States of America, **2** Ming Hsieh Department of Electrical and Computer Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, California, United States of America, **3** University of Sydney, Sydney, Australia, **4** Department of Animal Sciences, Purdue University, West Lafayette, Indiana, United States of America, **5** Department of Earth System Science, University of California, Irvine, California, United States of America, **6** Department of Electrical Engineering, University of Washington, Seattle, Washington, United States of America

* annethessen@gmail.com

Abstract

Decades of reductionist approaches in biology have achieved spectacular progress, but the proliferation of subdisciplines, each with its own technical and social practices regarding data, impedes the growth of the multidisciplinary and interdisciplinary approaches now needed to address pressing societal challenges. Data integration is key to a reintegrated biology able to address global issues such as climate change, biodiversity loss, and sustainable ecosystem management. We identify major challenges to data integration and present a vision for a "Data as a Service"-oriented architecture to promote reuse of data for discovery. The proposed architecture includes standards development, new tools and services, and strategies for career-development and sustainability.

Citation: Thessen AE, Bogdan P, Patterson DJ, Casey TM, Hinojo-Hinojo C, de Lange O, et al. (2021) From Reductionism to Reintegration: Solving society's most pressing problems requires building bridges between data types across the life sciences. PLoS Biol 19(3): e3001129. https://doi.org/10.1371/journal.pbio.3001129

Published: March 26, 2021

Funding: The National Science Foundation (NSF) nsf.gov supported the workshop where this paper was first developed. The National Institutes of Health Office of Research Infrastructure Programs (NIH/ORIP) orip.nih.gov funded the time of AET and MAH via an award to the Monarch Initiative (5 R24 OD011883). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

Introduction

Life on Earth is an interplay of interacting biological systems and geological processes that evolved over approximately 3 billion years and is represented by more than 2 million extant species. It is this complex system that creates the context for efforts to maintain global biodiversity while ensuring the health and well-being of our growing human population. Progress will require input from many disciplines to understand and manage our challenges [1]. Decades of reductionist research have led to extraordinary insights but have produced many subdiscipines with differing technical and social practices. If we are to solve societal problems, we must gain access to and bring together data from many disciplines. This is not straightforward because of the heterogeneity of data and associated conventions among communities. One clear and present challenge is how best to integrate data from the subdisciplines.

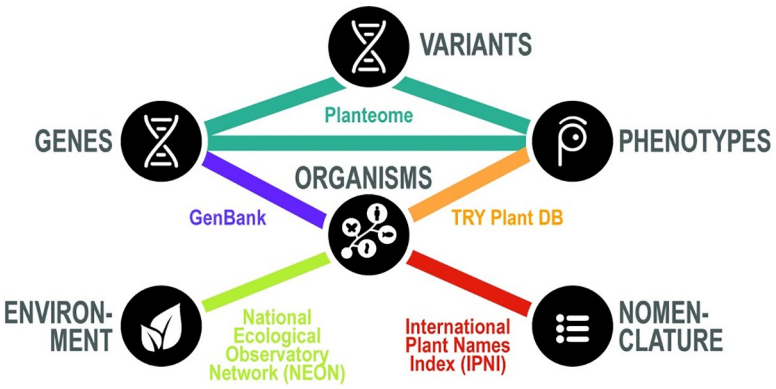

Fig 1. Reintegrating data to understand phenotype. Most biological data repositories only cover one part of the biological picture and must be integrated with other repositories in order to see the whole. Understanding phenotype requires data about genes, gene variants, organisms, environments, and taxonomy with nomenclature. Using plant phenotypes as an example, a minimum of 5 repositories are required to hold and curate relevant information. The repositories listed are only examples and do not represent all available resources.

https://doi.org/10.1371/journal.pbio.3001129.g001

Open access to data has the potential to democratize innovation by making it easier for third parties to reuse data and test solutions to complex problems that subdisciplines cannot address alone [2], but open access is merely a prerequisite. An important example of one of these complex problems is understanding the effect of genes and environments on observable phenotypes. Understanding phenotypes requires data about genes, variants, organisms, and environments, among others, and much of these data are open but not truly integrated (Fig 1). Understanding complex phenomena, such as expression of phenotypes, requires access to integrated data.

Our current limited ability to integrate data across scale, methodologies, and disciplines [3] impairs progress with multiscale, heterogeneous, non-Gaussian, and non-Markovian networks of dynamical systems [4]. Aligned with the vision of the National Science Foundation, we advocate for a comprehensive approach to data integration for predictive modelling of complex systems. Underlying issues of data sharing, integration, and reuse (Box 1) have been discussed widely [2,5–23]. Solutions to impediments have been proposed and variously implemented in the context of Data as a Service (DaaS). We use this term to point to evolving service-oriented infrastructures which provide data for third-party reuse. The infrastructure acquires data from primary sources and delivers fit-for-purpose content through trusted and curated repositories, inclusive of commercial and noncommercial agencies. DaaS aims not to serve a particular research agenda but is agile and adaptive such that it can support any project. The infrastructure is characterized by best practices, globally accepted standards, and is supported by a community of data experts, all of whom receive credit for their contributions [24]. One example of DaaS-oriented infrastructure for biology is CyVerse. Yet, challenges persist with increasing the scale and scope of data types that can be integrated and provided via DaaS, and with incentives for making data persistently available for use by third parties. Calls from high-profile scientific organizations [25,26] for unification of biological data are driven by improved computing power, new computational methods, maturing data standards, emerging exploratory protocols, advances in collaborative environments, changing attitudes about data sharing, and trustworthy data curation.

We advocate for a DaaS-informed strategy to build bridges between data types. Data-centered collaborations that are aware of the full scope of biology (Fig 2) will lead to novel cyberinfrastructures that enable cross-disciplinary data integration. A new cyberinfrastructure will

Box 1. Data integration challenges

Challenges in the nature of the data

- Data are highly variable;
- Data are collected on multiple spatiotemporal scales;
- Data generation has gaps;
- Data are not discoverable.

Challenges in the nature of biological systems

- Large biological systems are highly variable and dynamic;
- Biological systems do not comply with simple statistical models.

Challenges in the nature of data infrastructure

- The data infrastructure does not incentivize sharing;
- The data infrastructure is difficult to establish and sustain;
- Use of the data infrastructure requires specialized training;
- The data infrastructure may have restrictive licensing.

enable currently unimagined interdisciplinary investigations and promote paradigm shifts in biological research [27]. Building on previous reviews [28], we summarize outstanding barriers to effective data practices (Box 1) and make proposals that will help us overcome barriers.

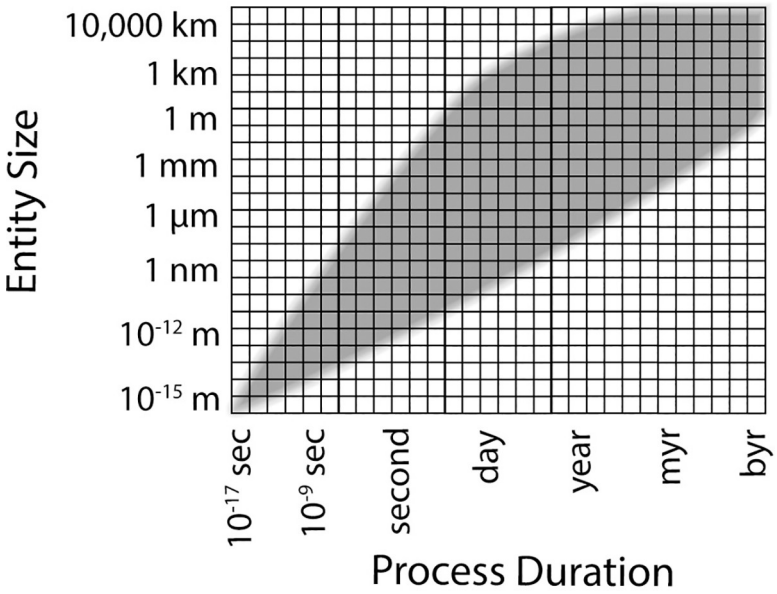

Fig 2. Envelope of life. Life sciences study entities (vertical axis) and processes (horizontal axis) that occur across a broad range of (logarithmic) scales. The shaded area emphasizes where biologically relevant processes occur.

https://doi.org/10.1371/journal.pbio.3001129.g002

Foundational infrastructure components

The development of a service-oriented DaaS architecture with appropriate human expertise and technical infrastructure will improve integration of currently separated data and enable transdisciplinary collaboration. The idea of access to data on demand across a broad front is not new. Several repositories and aggregators provide biological data on demand (e.g., [29–32]). We advocate for extending DaaS infrastructure to address persistent barriers to data sharing and integration. Below, we outline 7 challenges and propose opportunities to resolve each of them, which we refer to as foundational components.

Licensing of data and software

The open science movement rests on a foundation that research data, software code, and experimental methods should be publicly available and transparent unless privacy or confidentiality is an issue [8]. *The first foundational component of DaaS is straightforward, permissive, human- and machine-comprehensible licensing*. Licenses need to be simple to reduce confusion [33–36] and designed to allow automated access to data. A call to license data and software is not new, but licensing, copyright, and data use agreements are poorly understood [33,37,38], delaying their application. A restrictive license, in this context, is any license that places additional requirements or caveats on use of the data. Investment in data citation, data publication, microannotation, and nanopublication [39–44] will reduce the need for the restrictive licenses and nonstandard use agreements that are often in place to track reuse and impact. A global system of interconnected data with automated methods of tracking use and apportioning credit requires standardized, machine-readable licensing and data use agreements.

Data integration is still a largely manual task

As with people and ideas, data have been siloed within discipline-specific, project-specific, or lab-specific environments. The key to integrated data among silos is the universal adoption of community-developed standards [45] (Box 2). Even with most current standards, substantial manual effort can be required to map overlapping schemas or transform data from one context

Box 2. Standards development and governance

Successful standards require a sustained, iterative process of continued development that allows for changes to the standards, to metadata, and to the data they describe. Community-driven, consensus-building approaches, rather than an exclusively top-down approach, allow each community of practice to develop their own standards. This process is typically codified within a governance document that describes how to update the standard, resolve disputes, manage documents and versions, etc. Effective governance, including a Code of Conduct, can make a big difference in whether or not members feel welcome and effective, which drives participation. Governance should be well documented, community-driven, and reviewed at intervals that are sensible for the degree of change in the data and methods being standardized. The bottom-up development of sustainable, useful standards for data aggregation and integration necessitates a robust governance process that can represent community buy-in and provide a handbook for collaboration.

to another. Standards which describe data and metadata, document formats, content, protocols, and vocabularies are essential for automated integration [46]. With appropriate standards, the transformation workflow can be automated, it can include multiple quality checks, perform transformations, promote reproducibility, identify provenance, reduce manual errors and inconsistencies; all at a lower cost. The automation of integration will add machine-actionable links across repositories that will result in a network of data repositories. An example that integrates across repositories is Biolink, a graph-oriented data model that was developed for a biomedical use case but is being extended to the rest of biology [47]. Domain standards are a concrete way to increase interoperability and integration across repositories given that the basic elements (i.e., semantic types) of biology and the relationships between them are identified consistently across resources. Yet, the context-dependent nature of data transformation is not well represented by existing standards. A solution may lie with micro-schemas, highly localized data models to ensure accurate context-dependent data transformation similar to the GA4GH schemablocks concept [48] that can facilitate automation of highly contextual data like total cholesterol or ocean acidification.

The second foundational component of DaaS is standards to support machine actionable metadata and corresponding algorithms for automated standardization and integration.

Automated data integration requires standards. The incentive structures for most academics do not include the development of standards or scientific software as intellectual outputs. A lack of data standards impedes progress, and it is now timely to acknowledge efforts to improve standards and the tools and services which support their use [15,49].

Metadata are underappreciated

Despite the importance of metadata, their creation is still neglected [50]. Collecting metadata at the same time as data they describe is a recognized best practice [9], but this does not always happen and substantial amounts of data have been collected without standard metadata. *The third foundational component of DaaS is algorithms for automated metadata creation and standardization with documentation and provenance*. New tools are needed to automate the generation of metadata across data types and scales, where possible [46]. Machine learning (ML) and artificial intelligence (AI) can enhance metadata with standards and detect appropriate protocols for data normalization. High-priority automated tasks include named entity recognition, data and semantic typing, and protocol detection. Algorithms for semiautomated crowdsourced curation will benefit quality control and other tasks that cannot be fully automated [51]. Some entity recognition algorithms already exist [52–54] but have not yet received wide adoption because of problems with usability, sustainability, and discoverability of the tools; or because of the need of changes to work practices. Without a strong user community, it will be hard to recruit resources to create and improve these tools. One perception is that metadata preparation and documentation is altruistic and without significant impact [19,50]. While not a universal view [23,55], better professional rewards that value metadata creation and associated tools are needed.

The quality of data and metadata is variable

Issues relating to quality include social and technical aspects of the data, metadata, data providers, aggregators, and repositories [15,56–59]. Numerous studies explore trust in data sets and the expectations of users [15,42,43,60,61], but there is no widely adopted, formal process for judging data set quality. The peer review system for publications, even with its flaws [62], can provide a starting point for an assessment of the quality of data sets [42]. *The fourth foundational component of DaaS is a simple, predictable, and transparent system for peer review of*

data. While some repositories have a review process for submitted data sets that may include automated checks, and data publication journals can review data set documentation, routine rigorous peer review of data has not been implemented. One deterrent is that the pool of likely reviewers is already overburdened. If peer review of data sets is to have any hope of implementation, an infrastructure that puts reviews to good use and apportions credit for conducting reviews is needed. A supplementary approach is to use annotation technology to enable feedback on data sets and data atoms by end users [63].

Data use and contributions are hard to track

Researchers typically use citation metrics of publications as a measure of the impact of their career. The other types of activity, such contributions to an infrastructure, or third-party reuse of data, are often neglected because, historically, no comparable systems exist to track other endeavors. DataCite [40], an advocate for data access and discovery, developed a data set citation standard with DOI assignments that is used in several disciplines, but additional supporting infrastructure is needed to fully understand what a data citation, and the resulting metrics, means for the career of the producer and the value of the data themselves [64]. A roadmap for a system that supports standardized data-level metrics [65] was developed by MakeDataCount.org and is available for implementation. This roadmap fills many of the important technical gaps but requires resources to increase adoption across repositories, publishers, institutions, and researchers in order to create a system of data metrics comparable to publication metrics.

The mutable nature of data sets raises issues with identifiers and versioning that do not apply to publications [43]. For reproducibility, a published analysis must point to a persistent version of the data that were used, even if a small change was made after publication. In addition, credit needs to be apportioned appropriately when data sets are curated, subdivided, combined, vetted, and upgraded in ways that manuscripts are not. This raises several provenance and attribution issues that can only be addressed by well-documented versioning with a robust chain of provenance (for an example system, see [66]). When a data set is downloaded for analysis, that chain is usually broken, making it nearly impossible to communicate usage metrics back to the source and other agents in the data supply chain. Mechanisms and infrastructure that bring analyses to the data will better reveal the entire workflow so that it can be reproduced and refined. *The fifth foundational component of DaaS is a transparent pathway for preserving provenance and attribution within analytical environments*. Many computing environments for large data sets comply with this component because most researchers do not have the local resources to manipulate very large data sets. Researchers with small data sets that can be handled by spreadsheets are less likely to preserve these metadata. Collaborative environments with a support infrastructure similar to git, GitHub, or FilteredPush [63] engage all participants in data stewardship and to make the pathway of content flow, value-adding, and analysis visible.

Good data managers and curators are scarce

If we are to make full use of rapidly changing technology, we need data expertise coupled with in-depth biological knowledge [67]. People with such skills are rare. Increasingly, biologists will require data training, but this is not sufficient to create new advanced tools. Rather, we require a professional development structure for a community of biologists with advanced expertise in data management, curatorship, software development, and information science. *The sixth foundational component of DaaS is a method that attributes and credits the work of professional data managers and curators that is equal to manuscript citation and can accommodate microannotation and nanopublication*. The Contributor Attribution Model (CAM) [68]

used in combination with microannotation or nanopublication [39,69,70], wherein metadata are associated with individual data atoms (smallest usable elements), can underpin a system of attribution tracking where individual credit cascades through the long pathway of content flow [37,39]. CAM builds on the work of groups like CASRAI [71] by using CRediT [72] to inform the Contributor Role Ontology—the source of role terms in CAM [73]. (Domain-specific groups like CASRAI are an integral part of developing the community standards discussed above.) There are a few existing systems for recording attribution such as ORCID [74], OpenVIVO [75], and rescognito [76] that have begun to tackle the issue of credit for data work. With more transparency, the investment in making data more reusable becomes more measurable and removes the disincentive of working hard for no reward [19,22,60].

Sustainability for data remains elusive

The current strategy for funding scientific research leaves most data unsupported after completion of the project. An essential but often overlooked aspect of data integration is long-term preservation. Repositories, including museums and libraries, have the knowledge and expertise to provide sustainable preservation of data [27], but many data repositories accommodate only a single subdiscipline or data type (with a few exceptions, e.g., [29]). CoreTrustSeal promotes best practices for repositories committed to the long-term preservation and curation of data [77].

Fitness of data can corrode with time, and this requires maintenance of the schemas, metadata, and even data (Box 2). An example is when names of and concepts for taxa change [37,56], but their use as metadata remains uncorrected. While many repositories regularly update their content, including the creation of new data products (e.g., [78]), others lack the resources or disciplinary skills to make these updates quickly. This leads to dissatisfaction with the current ecosystem of long-term data support [15,56]. One reaction is for researchers to maintain data locally; but the probability that project-oriented data environments are available for reuse decreases by 17% per year [79]. *The seventh foundational component of DaaS is low-cost, reproducible workflows that convey data and metadata from creation to an accessible trusted repository that delivers data that are fit for purpose in perpetuity*. Lack of preservation resources places much of our collective digital knowledge in jeopardy, is dismissive of the investment in creating data, threatens future insights that might be drawn from data, and decreases our ability to engage in reproducible science.

Our vision of a reintegrating biology

It is inevitable that an extensive integrated data environment will emerge. With it will come new opportunities for discovery, devices to address problems with greater scale and scope, and the quality of insights will improve [80]. There are several leaders in this developing space, including CyVerse—an open science workspace for collaborative data-driven discovery that offers an infrastructure that supports data throughout its life cycle regardless of platform [81]. CyVerse supports access to high-performance computing and the packaging of data and software together in reusable "Research Objects" such as those used in the Whole Tale Environment [82]. A DaaS model can promote the emergence of a more extensive network of curated, interlinked data within an environment that is rich in tools and interoperable services. Progress is impeded because much of the required self-reinforcing infrastructure is absent. We emphasize 2 barriers to achieving the foundational components discussed here. First: motivating the sustained community participation that is needed to develop and implement discipline-specific data integration solutions—especially in respect of discipline-specific and domain-specific standards that enable the automated components that make large-scale

integration tractible. Second: the data citation and peer review infrastructure (beyond DOIs) needed to motivate professional participation in data-centric activities does not yet exist. The interconnected nature of these problems means that partial solutions, which are easier to fund, will not have the desired impact. The role of publishers and aggregators of intellectual output, like ORCID, in making this vision a reality cannot be overstated [83,84]. Some of the early progress with incentivizing data sharing were led by the requirements of publishers [85], and they remain a major driver of data sharing behaviors [22]. Publishers, researchers, and repositories will need to collaborate to adopt and enforce a standard of data citation and peer review that, combined with infrastructure supporting provenance, microattribution, annotation, and versioning proposed here, can perpetuate credit across a workflow. Well-formed attribution metadata can make essential data-related tasks just as visible as traditional publications. This is key to improving the academic incentive structure that currently demotivates investment in data-centric research practices.

Summary

Our ability to address issues that draw on many subdisciplines of biology will improve with integrated access to data across diverse types. Our vision is that disciplinary boundaries will break down to reveal a virtual pool of integrated data from many subdisciplines. This pool of data will need to be supported with an ecosystem of automated management processes, bridging metamodels, services, and semantic technologies. A DaaS approach can lead to decentralized repositories where knowledge and contributions are part of a distributed and shared global network that maintains provenance and attribution for all participants. To overcome impediments, we propose that the following 7 components will foster DaaS in biology:

- *Straightforward, permissive, human- and machine-comprehensible licensing;*
- *Standards to support machine actionable metadata;*
- *Algorithms that automate the creation of metadata where possible;*
- *A simple, predictable, and transparent system for peer review of data;*
- *A transparent pathway for preserving provenance and attribution within analytical environments;*
- *A method for attributing and crediting the work of data managers and curators;*
- *Low-cost, reproducible workflows that support the movement of data and metadata from creation to trusted repositories of data that are fit for purpose.*

Advances in automated data management practices, community standards, and data publication infrastructure put these components within reach. Investments in data infrastructure will increase data usability, impact, and marketability of data through a DaaS model and a shift in professional incentives that values investment in this area. Addressing these challenges will lead to an improved basis to answer current big questions in biology and contribute science-based solutions to the most pressing social and environmental problems.

Acknowledgments

The authors thank Nicole Vasilevksy, Ruth Duerr, and Chris Mungall for comments on the text and Julie McMurry for providing some of the figures.

References

1. Wolkovich EM, Regetz J, O'Connor MI. Advances in global change research require open science by individual researchers. Glob Chang Biol. 2012; 18:2102–2110.
2. Soranno PA, Cheruvelil KS, Elliott KC, Montgomery GM. It's Good to Share: Why Environmental Scientists' Ethics Are Out of Date. Bioscience. 2015; 65:69–73. https://doi.org/10.1093/biosci/biu169 PMID: 26955073
3. Thackeray SJ, Hampton SE. The case for research integration, from genomics to remote sensing, to understand biodiversity change and functional dynamics in the world's lakes. Glob Chang Biol. 2020; 26:3230–3240. https://doi.org/10.1111/gcb.15045 PMID: 32077186
4. Bogdan P. Taming the Unknown Unknowns in Complex Systems: Challenges and Opportunities for Modeling, Analysis and Control of Complex (Biological) Collectives. Front Physiol. 2019; 10:1452. https://doi.org/10.3389/fphys.2019.01452 PMID: 31849703
5. Thessen AE, Fertig B, Jarvis JC, Rhodes AC. Data Infrastructures for Estuarine and Coastal Ecological Syntheses. Estuaries Coast. 2016; 39:295–310.
6. Thessen AE, Patterson DJ. Data issues in the life sciences. Zookeys. 2011; 150. https://doi.org/10.3897/zookeys.150.1766 PMID: 22207805
7. Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, Crosas M, et al. Ten simple rules for the care and feeding of scientific data. PLoS Comput Biol. 2014; 10:e1003542. https://doi.org/10.1371/journal.pcbi.1003542 PMID: 24763340
8. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016; 3:160018. https://doi.org/10.1038/sdata.2016.18 PMID: 26978244
9. White EP, Baldridge E, Brym ZT, Locey KJ, McGlinn DJ, Supp SR. Nine simple ways to make it easier to (re) use your data. Ideas Ecol Evol. 2013; 6. Available from: https://ojs.library.queensu.ca/index.php/IEE/article/view/4608
10. Leonelli S. The challenges of big data biology. Elife. 2019; 8. https://doi.org/10.7554/eLife.47381 PMID: 30950793
11. Data sharing and the future of science. Nat Commun. 2018; 9:2817. https://doi.org/10.1038/s41467-018-05227-z PMID: 30026584
12. Enke N, Thessen A, Bach K, Bendix J, Seeger B, Gemeinholzer B. The user's view on biodiversity data sharing: Investigating facts of acceptance and requirements to realize a sustainable use of research data. Eco Inform. 2012; 11:25–33.
13. Gemeinholzer B, Vences M, Beszteri B, Bruy T, Felden J, Kostadinov I, et al. Data storage and data re-use in taxonomy—the need for improved storage and accessibility of heterogeneous data. Org Divers Evol. 2020; 20:1–8.
14. König C, Weigelt P, Schrader J, Taylor A, Kattge J, Kreft H. Biodiversity data integration-the significance of data resolution and domain. PLoS Biol. 2019; 17:e3000183. https://doi.org/10.1371/journal.pbio.3000183 PMID: 30883539
15. Tenopir C, Rice NM, Allard S, Baird L, Borycz J, Christian L, et al. Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. PLoS ONE. 2020; 15:e0229003. https://doi.org/10.1371/journal.pone.0229003 PMID: 32160189
16. Qin J, Ball A, Greenberg J. Functional and architectural requirements for metadata: supporting discovery and management of scientific data. International Conference on Dublin Core and Metadata Applications. dcpapers.dublincore.org; 2012. pp. 62–71.
17. Zimmerman AS. New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data. Sci Technol Human Values. 2008; 33:631–652.
18. Faniel IM, Zimmerman A. Beyond the Data Deluge: A Research Agenda for Large-Scale Data Sharing and Reuse. Int J Digit Curation. 2011; 6:58–69.
19. Pronk TE, Wiersma PH, Weerden A van, Schieving F. A game theoretic analysis of research data sharing. PeerJ. 2015; 3:e1242. https://doi.org/10.7717/peerj.1242 PMID: 26401453
20. Savage CJ, Vickers AJ. Empirical study of data sharing by authors publishing in PLoS journals. PLoS ONE. 2009; 4:e7078. https://doi.org/10.1371/journal.pone.0007078 PMID: 19763261
21. Chawinga WD, Zinn S. Global perspectives of research data sharing: A systematic literature review. Libr Inf Sci Res. 2019; 41:109–122.
22. Kim Y, Stanton JM. Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis: Institutional and Individual Factors Affecting Scientists' Data Sharing Behaviors: A Multilevel Analysis. J Assn Inf Sci Tec. 2016; 67:776–799.

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

24. Rouse M. What is Data as a Service (DaaS)? In: SearchDataManagement [Internet]. TechTarget; 16 May 2019 [cited 2020 Apr 2]. Available from: https://searchdatamanagement.techtarget.com/definition/data-as-a-service

25. Shorthouse DP, Patterson D, Stenseth NC. Unifying Biology Through Informatics (UBTI) a new programme of the International Union of Biological Sciences. BISS. 2017; 1:e20431.

26. ESA. Moving Forward with Ecological Informatics and Reproducibility. In: EcoTone: News and Views on Ecological Science [Internet]. [cited 2020 May 26]. Available from: https://www.esa.org/esablog/research/moving-forward-with-ecological-informatics-and-reproducibility/

27. McMurry JA, Köhler S, Washington NL, Balhoff JP, Borromeo C, Brush M, et al. Navigating the Phenotype Frontier: The Monarch Initiative. Genetics. 2016; 203:1491–1495. https://doi.org/10.1534/genetics.116.188870 PMID: 27516611

28. Renaut S, Budden AE, Gravel D, Poisot T, Peres-Neto P. Management, Archiving, and Sharing for Biologists and the Role of Research Institutions in the Technology-Oriented Age. Bioscience. 2018; 68:400–411.

29. Vision T. The Dryad Digital Repository: Published evolutionary data as part of the greater data ecosystem. Nature Precedings. 2010. https://doi.org/10.1038/npre.2010.4595.1

30. CyVerse Home. [cited 2020 Oct 9]. Available from: https://cyverse.org/

31. Telenius A. Biodiversity information goes public: GBIF at your service. Nord J Bot. 2011; 29:378–381.

32. Michener W, Vieglais D, Vision T, Kunze J, Cruse P, Janée G. DataONE: Data Observation Network for Earth—Preserving data and enabling innovation in the biological and environmental sciences. D-Lib Magazine. 2011; 17:12.

33. Hagedorn G, Mietchen D, Morris RA, Agosti D, Penev L, Berendsohn WG, et al. Creative Commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information. Zookeys. 2011;127–149. https://doi.org/10.3897/zookeys.150.2189 PMID: 22207810

34. Carbon S, Champieux R, McMurry JA, Winfree L, Wyatt LR, Haendel MA. An analysis and metric of reusable data licensing practices for biomedical resources. PLoS ONE. 2019; 14: e0213090. https://doi.org/10.1371/journal.pone.0213090 PMID: 30917137

35. Oxenham S. Legal maze threatens to slow data science. Nature. 2016; 536:16–17. https://doi.org/10.1038/536016a PMID: 27488781

36. Analyzing the licenses of all 11,000+ GBIF registered datasets—Peter Desmet. [cited 2020 Mar 31]. Available from: http://peterdesmet.com/posts/analyzing-gbif-data-licenses.html

37. Patterson DJ, Egloff W, Agosti D, Eades D, Franz N, Hagedorn G, et al. Scientific names of organisms: attribution, rights, and licensing. BMC Res Notes. 2014; 7:79. https://doi.org/10.1186/1756-0500-7-79 PMID: 24495358

38. Egloff W, Agosti D, Kishor P, Patterson D, Miller J. Copyright and the Use of Images as Biodiversity Data. Riogrande Odontol. 2017; 3:e12502.

39. Patrinos GP, Cooper DN, van Mulligen E, Gkantouna V, Tzimas G, Tatum Z, et al. Microattribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain. Hum Mutat. 2012; 33:1503–1512. https://doi.org/10.1002/humu.22144 PMID: 22736453

40. DataCite. Welcome to DataCite. 2018. Available from: https://datacite.org/

41. Mooney H. A Practical Approach to Data Citation: The Special Interest Group on Data Citation and Development of the Quick Guide to Data Citation. IASSIST Quarterly. 2014. p. 71. https://doi.org/10.29173/iq240

42. Kratz J, Strasser C. Data publication consensus and controversies. F1000Res. 2014; 3:94. https://doi.org/10.12688/f1000research.3979.3 PMID: 25075301

43. Parsons MA, Duerr RE, Jones MB. The History and Future of Data Citation in Practice. Data Sci J. 2019; 18. Available from: https://datascience.codata.org/articles/10.5334/dsj-2019-052/print/ PMID: 31579260

44. Tang YA, Pichler K, Füllgrabe A, Lomax J, Malone J, Munoz-Torres MC, et al. Ten quick tips for biocuration. PLoS Comput Biol. 2019; 15:e1006906. https://doi.org/10.1371/journal.pcbi.1006906 PMID: 31048830

45. Lapatas V, Stefanidakis M, Jimenez RC, Via A, Schneider MV. Data integration in biological research: an overview. J Biol Res. 2015; 22:9. https://doi.org/10.1186/s40709-015-0032-5 PMID: 26336651

46. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. Inf Fusion. 2019; 50:71–91. https://doi.org/10.1016/j.inffus.2018.09.012 PMID: 30467459

47. biolink-model. Github; Available from: https://github.com/biolink/biolink-model
48. ga4gh-schemablocks.github.io. [cited 2020 Nov 13]. Available from: https://schemablocks.org/
49. Poisot T, Bruneau A, Gonzalez A, Gravel D, Peres-Neto P. Ecological Data Should Not Be So Hard to Find and Reuse. Trends Ecol Evol. 2019; 34:494–496. https://doi.org/10.1016/j.tree.2019.04.005 PMID: 31056219
50. Edwards PN, Mayernik MS, Batcheller AL, Bowker GC, Borgman CL. Science friction: data, metadata, and collaboration. Soc Stud Sci. 2011; 41:667–690. https://doi.org/10.1177/0306312711413314 PMID: 22164720
51. Lock A, Harris MA, Rutherford K, Hayles J, Wood V. Community curation in PomBase: enabling fission yeast experts to provide detailed, standardized, sharable annotation from research publications. Database. 2020; 2020. https://doi.org/10.1093/database/baaa028 PMID: 32353878
52. Mozzherin D, Myltsev AA, Patterson D. Finding scientific names in Biodiversity Heritage Library, or how to shrink Big Data. BISS. 2019; 3:e35353.
53. Furrer L, Jancso A, Colic N, Rinaldi F. OGER++: hybrid multi-type entity recognition. J Chem. 2019; 11:7. https://doi.org/10.1186/s13321-018-0326-3 PMID: 30666476
54. Gonçalves RS, O'Connor MJ, Martínez-Romero M, Egyedi AL, Willrett D, Graybeal J, et al. The CEDAR Workbench: An Ontology-Assisted Environment for Authoring Metadata that Describe Scientific Experiments. Semant Web ISWC. 2017; 10588:103–110. https://doi.org/10.1007/978-3-319-68204-4_10 PMID: 32219223
55. Pronk TE. The Time Efficiency Gain in Sharing and Reuse of Research Data. Data Sci J. 2019. Available from: https://datascience.codata.org/article/10.5334/dsj-2019-010/ PMID: 31579260
56. Franz NM, Sterner BW. To increase trust, change the social design behind aggregated biodiversity data. Database. 2018; 2018. https://doi.org/10.1093/database/bax100 PMID: 29315357
57. Yoon A. Data reusers' trust development. J Assoc Inf Sci Technol. 2017; 68:946–956.
58. Belbin L, Daly J, Hirsch T, Hobern D, Salle JL. A specialist's audit of aggregated occurrence records: An "aggregator"s' perspective. Zookeys. 2013;67–76.
59. Mesibov R. A specialist's audit of aggregated occurrence records. Zookeys. 2013;1–18.
60. Kratz JE, Strasser C. Researcher perspectives on publication and peer review of data. PLoS ONE. 2015; 10:e0117619. https://doi.org/10.1371/journal.pone.0117619 PMID: 25706992
61. Parsons MA, Duerr R, Minster J-B. Data Citation and Peer Review. Eos Trans AGU. 2010; 91:297.
62. Hausmann L, Murphy SP, Publication Committee of the International Society for Neurochemistry (ISN). The challenges for scientific publishing, 60 years on. J Neurochem. 2016; 139 Suppl 2:280–287.
63. Morris RA, Dou L, Hanken J, Kelly M, Lowery DB, Ludäscher B, et al. Semantic annotation of mutable data. PLoS ONE. 2013; 8:e76093. https://doi.org/10.1371/journal.pone.0076093 PMID: 24223697
64. Robinson-Garcia N, Mongeon P, Jeng W, Costas R. DataCite as a novel bibliometric source: Coverage, strengths and limitations. J Informet. 2017; 11:841–854.
65. Pesch O. COUNTER: Looking Ahead to Release 5 of the COUNTER Code of Practice. Ser Libr. 2016; 71:83–90.
66. Missier P. Data trajectories: tracking reuse of published data for transitive credit attribution. Int J Digit Curation. 2016; 11:1–16.
67. Markowetz F. All biology is computational biology. PLoS Biol. 2017; 15:e2002050. https://doi.org/10.1371/journal.pbio.2002050 PMID: 28278152
68. Welcome to the Contributor Attribution Model—Contributor Attribution Model documentation. [cited 2020 May 31]. Available from: https://contributor-attribution-model.readthedocs.io/en/latest/
69. Raciti D, Yook K, Harris TW, Schedl T, Sternberg PW. Micropublication: incentivizing community curation and placing unpublished data into the public domain. Database. 2018;2018. https://doi.org/10.1093/database/bay013 PMID: 29688367
70. Kuhn T, Meroño-Peñuela A, Malic A, Poelen JH, Hurlbert AH, Centeno Ortiz E, et al. Nanopublications: A Growing Resource of Provenance-Centric Scientific Linked Data. 2018 IEEE 14th International Conference on e-Science (e-Science). ieeexplore.ieee.org; 2018. pp. 83–92.
71. Welcome to CASRAI. 6 Oct 2019 [cited 2021 Jan 27]. Available from: https://casrai.org/
72. Holcombe AO. Contributorship, Not Authorship: Use CRediT to Indicate Who Did What. Publications. 2019; 7:48.
73. Vasilevsky NA, Hosseini M, Teplitzky S, Ilik V, Mohammadi E, Schneider J, et al. Is authorship sufficient for today's collaborative research? A call for contributor roles. Account Res. 2021; 28:23–43. https://doi.org/10.1080/08989621.2020.1779591 PMID: 32602379

74. Haak LL, Fenner M, Paglione L, Pentz E, Ratner H. ORCID: a system to uniquely identify researchers. Learn Publ. 2012; 25:259–264.

75. Ilik V, Conlon M, Triggs G, White M, Javed M, Brush M, et al. OpenVIVO: Transparency in Scholarship. Front Res Metr Anal. 2018; 2:12.

76. Wynne R. Got a DOI? Claim and Give Some CRediT! 2019. https://doi.org/10.6084/m9.figshare.9733595.v1

77. Dillo I, de Leeuw L. CoreTrustSeal. Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen & Bibliothekare. 2018; 71:162–170.

78. Baker KS, Duerr RE, Parsons MA. Scientific knowledge mobilization: Co-evolution of data products and designated communities. Int J Digit Curation. 2016; 10:110–135.

79. Vines TH, Albert AYK, Andrew RL, Débarre F, Bock DG, Franklin MT, et al. The availability of research data declines rapidly with article age. Curr Biol. 2014; 24:94–97. https://doi.org/10.1016/j.cub.2013.11.014 PMID: 24361065

80. Molloy JC. The Open Knowledge Foundation: open data means better science. PLoS Biol. 2011; 9: e1001195. https://doi.org/10.1371/journal.pbio.1001195 PMID: 22162946

81. Swetnam TL, Walls R, Devisetty UK, Merchant N. CyVerse: a Ten-year Perspective on Cyberinfrastructure Development, Collaboration, and Community Building. AGUFM. 2018. p. IN23B–0767.

82. Brinckman A, Chard K, Gaffney N, Hategan M, Jones MB, Kowalik K, et al. Computing environments for reproducibility: Capturing the "Whole Tale." Future Gener Comput Syst. 2019; 94:854–867.

83. Lin J, Strasser C. Recommendations for the role of publishers in access to data. PLoS Biol. 2014; 12: e1001975. https://doi.org/10.1371/journal.pbio.1001975 PMID: 25350642

84. Cousijn H, Kenall A, Ganley E, Harrison M, Kernohan D, Lemberger T, et al. A data citation roadmap for scientific publishers. Sci Data. 2018; 5:180259. https://doi.org/10.1038/sdata.2018.259 PMID: 30457573

85. Strasser BJ. The experimenter's museum: GenBank, natural history, and the moral economies of biomedicine. Isis. 2011; 102:60–96. https://doi.org/10.1086/658657 PMID: 21667776