Title:

Reintegrating biology: A data-centric approach.

Authors:

Anne Thessen*, Paul Bogdan, David J. Patterson, Theresa Casey, Cesar Hinojo, Orlando de Lange, Melissa A. Haendel

Abstract

Decades of reductionist scientific approaches have led to spectacular progress and the proliferation of biological sub-disciplines, each spawning its own technical and social practices regarding data. This fragmented landscape poses substantial hurdles to the multi-disciplinary approaches needed to address pressing societal challenges. Data integration is key to the reintegration of biology and the pursuit of global questions such as climate change, biodiversity loss, and sustainable ecosystem management. Here, we define the primary challenges in data integration and present a vision for a Data as a Service (DaaS) oriented architecture that enables frictionless data reuse, hypothesis testing, and discovery. The proposed data integration infrastructure includes standards development, a suite of tools and services, and strategies for education and sustainability.

Introduction

Life on Earth is an interplay of interacting biological systems and geological processes that evolved over ~ 3 billion years and is represented by more than 2 million extant species. Maintaining global biodiversity while ensuring the health and well-being of our growing human population will require experts from across diverse disciplines to draw from all classes of information [1,2] in order to integrate data to understand and solve complex challenges. Decades of reductionist research, while leading to extraordinary insights, have created technical and social silos around the very disciplines that need to unite to solve societal problems. As a result, the field of biology has fractured into thousands of subdisciplines, each operating within its own research culture. The reintegration of subdisciplines can be achieved through the reintegration of data; but due to heterogeneity of both data and communities, this is a serious and pressing challenge.

We propose that a data-centric integration approach that focuses on building bridges between data types (in addition to human-focused communication) will more successfully reintegrate biology. Making data openly accessible is a prerequisite for broad data integration, and the potential as well as the associated problems of open data have been widely discussed [3–17]. 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 sub-disciplines cannot address alone [18]. 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). Making data open is merely a prerequisite. There are a host of other problems that must be addressed to evolve from a system of distributed data to integrated data [19]. In this paper, we highlight what we consider to be ten key challenges to biological data integration beyond making data open. Building on previous reviews [20], we discuss how solving these problems will create effective data practices and a Data as a Service (DaaS) oriented architecture for data integration in the life sciences.

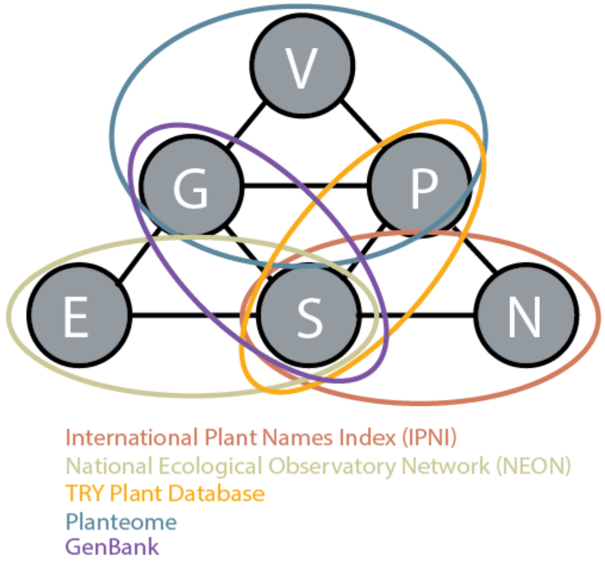
V
G
P
E
S
N
International Plant Names Index (IPNI)
National Ecological Observatory Network (NEON)
TRY Plant Database
Planteome
GenBank

Figure 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 (P) requires data about genes (G), gene variants (V), organisms (S), environments (E), and taxonomy with nomenclature (N). Using plant phenotypes as an example, a minimum of five repositories are required to hold and curate relevant information. The repositories listed are only examples and do not represent all available resources.

Background

Current Landscape

Reductionist approaches have created a data landscape that is rich, but ill-suited to study phenomena that extend across broad temporal and physical scales (Fig. 2). Simple mathematical models are not designed to describe the large-scale interacting processes that exist in biological systems [21,22]. Aligned with the vision of the National Science Foundation, we advocate an expansive research approach that involves data integration and modelling for enhanced understanding of complex systems and the framing of hypotheses that could never be conceived of in the absence of a systems-level view. Answers to complex questions require data integration, itself not a simple issue [10,23–25], which goes beyond making data available as distributed data sets or as aggregated data resources. The need for unification through informatics has been recognized by the International Union of Biological Sciences [26], the American Medical Informatics Association [27], and the Ecological Society of America [28], among many others. We categorize data integration challenges in three classes: the ‘nature of data’, the ‘nature of biological systems,’ and the ‘nature of data infrastructure’. Progress is now possible because of improvements in computing power, computational methods, maturing data standards, exploratory protocols, and attitudes about data sharing.

Challenges in the Nature of the Data

Data are highly variable. Integration is challenged by the myriad types of data that range from discrete variables (e.g., presence or absence of an anatomical structure), to continuous values (e.g., protein concentrations), graph-like structures (e.g., 3D chromatin conformation), and compound objects (e.g., images). The same data type can appear using different formats and terminologies, and alongside sparse or variable metadata. Multiple, differentially-annotated copies of the same data may be replicated across repositories, leading to erroneous duplication within an integrated data set. Even mathematical characteristics of the data, such as variability and noise, can be inconsistent. As a result, data are highly heterogeneous in form, terminology, and scale (for example, [2,25,29,30]). Thus, data integration should be designed “fit for purpose,” requiring special consideration of quality and fitness as well as user need to avoid erroneous results from variable data combined improperly [31,32].

Data are collected on multiple spatiotemporal scales. Biological processes occur over a wide range of temporal (femtoseconds to billions of years) and spatial (subatomic particles to the entire biosphere) scales (Fig. 2). Integrating data relating to complex phenomena (such as the role of ocean biology in carbon sequestration) requires combining data about molecular processes (photosynthesis) with data collected at the global scale (remote sensing). Even measurements of the same process can be a challenge to integrate if taken at different frequencies. Methods for integrating data with scale mismatches are in an early stage of development and tend to be very application specific [33,34]. The spatial and temporal contexts in which data were collected impacts integration methodology, but no best-practices have been developed and the context is often lost.

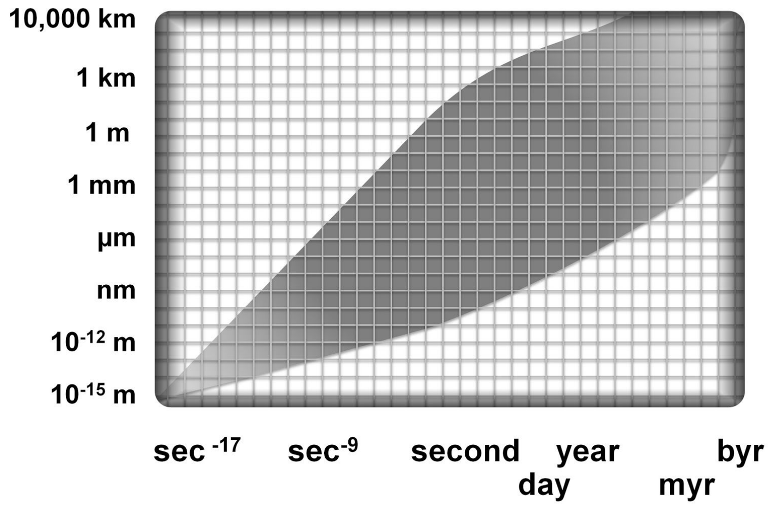

Figure 2: Envelope of Life. Life sciences study entities (vertical axis) and processes (horizontal axis) that occur across a broad range of (logarithmic) scales. The shaded area emphasizes where biologically-relevant processes occur.

Data generation has gaps. The biosphere is very unevenly sampled [35] and this has consequences for data integration [2]. Sampling can be biased by scope of study or ease of measurement, resulting in data with poor representation of variation across time, space, and biological levels. There is also difficulty in uniformly sensing variables across large spatial dimensions, which can lead to a granularity “mismatch” between data sets. Sampling bias, whether implicit, explicit, or caused by the limitations of the instruments, can skew the perceived importance of a factor in a system, and more heavily studied systems will have undue influence if this is not controlled for in an analysis. A landscape analysis that inventories available data sets and recognizes their limitations can reveal gaps *a priori* so they can be appropriately handled. A failure to acknowledge these gaps in an integrated data set could lead to flawed insights.

Data are not discoverable. The first step in building an integrated data set is finding all relevant data. Impediments to data discovery lie, in part, within discipline-specific data cultures [3,14]. These cultures influence the use of standards and identifiers that make data discoverable, the use of repositories, and the selection of licenses to make data accessible. The fields of meteorology, economics, and astronomy have been proactive in making data discoverable and shareable, and these efforts are reflected in the robust mechanisms available for querying datasets, as well as the predictive models developed [36–38]. On the other hand, data repositories and sharing in the biological sciences is less mature and fraught with social and technical barriers [9,10,39,40]. The library community has developed discipline-agnostic metadata standards for discovery, but some specialized data sets can have details that are difficult to represent in general research standards [41,42]. Moreover, people increasingly rely on computational agents (such as an internet browser) for data discovery, but without pervasive and consistent use of identifiers [43], data standards [44], metadata standards [4], and controlled vocabularies [45] these search tools are not fully effective, meaning that word-of-mouth and citations in publications are still essential to data discovery [46–48]. Disciplines that rely on pre-digital data sources, such as taxonomy, have large gaps in online content [49]. Thus, to advance towards data integration in the biological sciences, the culture around data sharing and discovery must move beyond reliance on word-of-mouth to establish universal standards of metadata quality and data preservation and support a workflow for digitizing legacy information.

Challenges in the Nature of Biological Systems

Large biological systems are highly variable and dynamic. The natural systems which biological data represent are filled with feedback loops, trajectories, stochasticity, memory, and emergent properties, and many are rarely or never in steady-state equilibrium. They are characterized by multiple concurrent processes that may interact in complicated ways. Biological systems can change their model structure or parameters over time [50,51]. Analysis of DNA sequences, [52,53] gene expression, [51] heart rate, [54,55], primary production [56], and brain activity [57,58] demonstrate that the current state of a system depends not only on what is happening now, but also what happened in the past [55,59–62]. Integrating data from a specific time point is not enough to understand the system. Data about what happened the day, week, month, year, or century before or in prior organismal generations may be needed for an exact understanding [63]. In addition to these known sources of variability, there are likely several unknown sources of variability which are not recognized or understood. For instance, the dynamics between genes and linked transcription factors in gene regulatory networks in *Escherichia coli* and *Saccharomyces cerevisiae* exhibit rich variability, multifractal and long-range, cross-correlated behavior, yet the sources and functional implications are not known [50,51]. Similar arguments apply to the observed rich variability and multifractality in healthy heart rate [54], blood glucose [64], and brain activity [57,58] time series. Accounting for dynamic change in analysis and interpretation imposes challenges on its own, and thus further challenges the integration of data across disciplines and scales.

Biological systems do not comply with simple statistical models. Since measuring everything is not possible, scientists measure a subsample and apply the knowledge gained to the whole using statistics to communicate boundaries. This practice is standard in research, but the dynamic nature of living systems makes it difficult to know just how far these generalizations are valid. Assuming that a collection of random samples of a process must represent the statistical properties of the average of the entire process (e.g., ergodicity) often falls short, limiting our ability to generalize about the whole based on a subsample. In addition, the assumption that data collected from one organism can apply to another organism if they are the same species is common; however, organisms will differ based on genetic makeup and environmental context, age, gender, and history (inter alia), creating a degree of variability between organisms of the same species [65–69]. Knowing exactly how far to carry this generalization is difficult because, while individual organisms may be regarded as discrete "things", species and other taxa are evolutionary processes that are continually changing. The science of discovering and defining species, taxonomy, involves making hypotheses about discontinuities among these processes, communicated via a taxonomic name or other label and through classification systems. While taxonomic names are a very useful class of metadata [70], we have noted that their effective use must overcome difficulties, such as changes in the nomenclature and classification over time [71] and their unique identification, among others. The dynamic and fluid nature of taxa makes it impossible to establish unambiguous criteria for what makes a taxon and which organisms belong to a specific taxon [72], eroding the certainty in this form of generalization as we increase the scale of integration for analyses. Statistical assumptions such as these are an important part of research, but are challenged by the complexity of biological systems.

Challenges in the Nature of the Data Infrastructure

Data infrastructure does not incentivize sharing. An important barrier to large-scale biological data integration is difficulty in getting data from individual researchers [73]. The vast majority of research data are stored locally and not preserved for reuse [74,75]. Despite demonstrated benefits [5,6], data sharing is viewed as largely altruistic with little professional reward [11]. Even the Research Parasite Award that honors outstanding secondary analysis is bestowed on the data consumers rather than the providers [76,77]. Some progress has been made in the form of data journals, data citation guidelines [78–80], and the acceptance of data products as valued research output in some contexts. The genomics community has accepted sharing as a societal norm by depositing sequence information with members of the International Nucleotide Sequence Database Collaboration, such as GenBank, and gene expression data into Gene Expression Omnibus. Currently, sharing is driven largely by requirements from funding agencies and publishers [16,81,82]. However, these organizations currently do little to ensure adequate or quality data sharing. For example, reviewers are given little guidance or opportunity to review the data sharing or management plan and there is no follow-up to ensure that the plan was indeed followed. The mismatch between the large number of researchers who claim to be willing to share data and those who actually make their data easily available for reuse suggests that sharing would increase substantially if the proper infrastructure were in place [9,75,83–89]. Structures to reward good data sharing practice, including attribution or co-authorship, would undoubtedly increase the volume and quality of shared datasets [16,73,81]. Currently, the benefits of reusable data fall mostly to the consumer.

Data infrastructure is difficult to sustain. Large scale data integration requires significant infrastructure investments for preservation and discovery, sometimes years before the benefits are apparent or known. Principal Investigators are incentivized to prioritize one-time dissemination of findings as peer-reviewed publications, rather than investing in long-term, sustainable data management infrastructure and practices that reward for reuse; thus, researchers are not motivated or prepared to support data over the long-term [16,73,90]. This includes everything from not allocating funds for data maintenance and dissemination to not valuing it when requested during grant reviews. Currently, the only available model for researchers to preserve their data for reuse is to transfer the data to a trusted repository, but many online databases are fragile and have low persistence [91]. Most are designed to support the very niche scientific communities and are not readily interoperable nor are the data transferrable if needed. It can be difficult to make the case for financial support for data repositories because the costs are immediate, but the benefits are long-term and future technologies uncertain. The long time horizons, diffuse stakeholders, and misaligned incentives make defining value propositions and who will pay difficult [92].

Data infrastructure use requires specialized training. Data infrastructure, even if well designed, will require a degree of data-literacy from its users, the absence of which will hinder participation [73,93]. Data science courses have only recently begun to be offered at universities, often parallel to and not integrated into traditional courses, meaning that many biologists continue to receive inadequate training in informatics methods. Unfortunately, almost no content is offered on data standards and best practices for management and sharing. Informal education, such as the Carpentries [94] (a global community of instructors teaching basic programming and data science skills) and online tutorials fill some of the training gap. Nevertheless, the expertise required to support data integration using varied and complicated data models and web services continues to exceed what most researchers are willing and able to learn. A higher standard of data literacy for all biologists generating and disseminating data coupled with a new cadre of well supported data professionals is necessary to integrate data at scale.

Data infrastructure can include restrictive licensing. Even theoretically open data can be constrained by a quagmire of legal uncertainty due to complicated, missing, or non-standard licensing and data use agreements (Fig. 3). Such complications require significant time for a data consumer to address [95,96], and for an integrated data set, managing the conflicting assortment of licenses and data use agreements can be nearly impossible [97]. Many data providers are just as uncertain about licensing their data and are not aware of the full legal implications of different licenses [98,99]. While not traditionally thought of as infrastructure, licensing and usage agreements can support or hinder data integration and reuse just as much as software or hardware and require investment to develop and maintain.

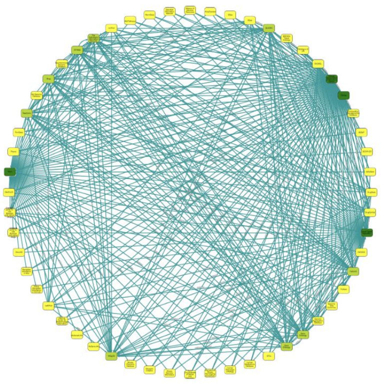

Figure 3: Effect of licensing on data integration. A data license can have a huge impact on the integration potential (edges) of individual data sets (nodes). Data sets with more permissible licenses (darker nodes) can be used far more than those with more restrictive licenses (lighter nodes). Of the 64 data sources studied by the Reusable Data Project [100], 34% used a non-standard license [99]. The data sets with the most permissive license could be used by all 63 other data sources while the data set with the most restrictive license could not be used by any other data source in the project.

Foundational Infrastructure Components

Developing solutions to these data challenges will require a consistent and dedicated effort to develop new technologies and community research norms. To this end, we advocate for the development of a service-oriented architecture for Data as a Service (DaaS) with human and technical infrastructure. The overall goal of DaaS is to support submission and reuse of data for frictionless integration across diverse data types and disciplines. The idea of DaaS is not new, and several repositories and aggregators provide biological data to a user on demand with varying numbers of users. We advocate for an expansion of the DaaS philosophy to include a full range of data types and expansion of the DaaS infrastructure to address the social barriers to data sharing. Below, we propose seven foundational components of DaaS, discuss existing elements, and what needs to be developed or expanded to enable the realization of the overall goal.

Open access to data wherever possible. The call to make research data, software code, and experimental methods publicly available and transparent is coming from within the fields of biology and is required by many funding sources (e.g., the NSF data management plan and NIH data and resource sharing plan). Advocates of making data open say it is the only way to address the lack of reproducibility in scientific findings and the best way for researchers to gather the range of observations needed to increase the rate of discovery and identify large-scale trends [39]. Data sharing can democratize access to data types that require expensive equipment which improves access for researchers at small institutions. A robust culture of data sharing has the potential to revolutionize the social aspect of research. *We recommend making data as open as possible using computable formats and permissive, standard licensing, with appropriate restrictions for confidentiality to protect privacy and sensitive species.* This recommendation parallels the call to make data Findable, Accessible, Interoperable and Reusable (FAIR) [101], but we also add that there should be testable metrics for successful reusable data sharing [102]. While there have been massive advances in making data available, open data are only the first step in frictionless reuse and not all sub-disciplines have - or can have - equally open data. In such cases, we recommend approaches that provision subsets of the data or synthetic derivatives, such that the data can be found and computed upon even in the absence of ability to share data in its fullest (for example Personal Health Information or data with national security issues).

Data-centered collaboration. In addition to people and ideas being siloed into biological sub-disciplines, data have been separated as well. As a result, no single discipline or data repository has the full biological picture or "model" (Fig 4A). Data-centered collaborations (i.e., collaborations which focus on bringing data together AND on bringing people together) need to be prioritized and incentivized, as they will create specific use cases for designing infrastructures that support cross-disciplinary data integration (Fig 4). Once this infrastructure is established, data can be shared and integrated to answer a variety of questions that transcend disciplines (Fig 4C). The shifting emphasis from supporting collaborations between individual people to supporting collaborations between data types and sources will add bridges across existing information architectures rather than creating completely new architectures, thereby reducing discipline-specific data silos.

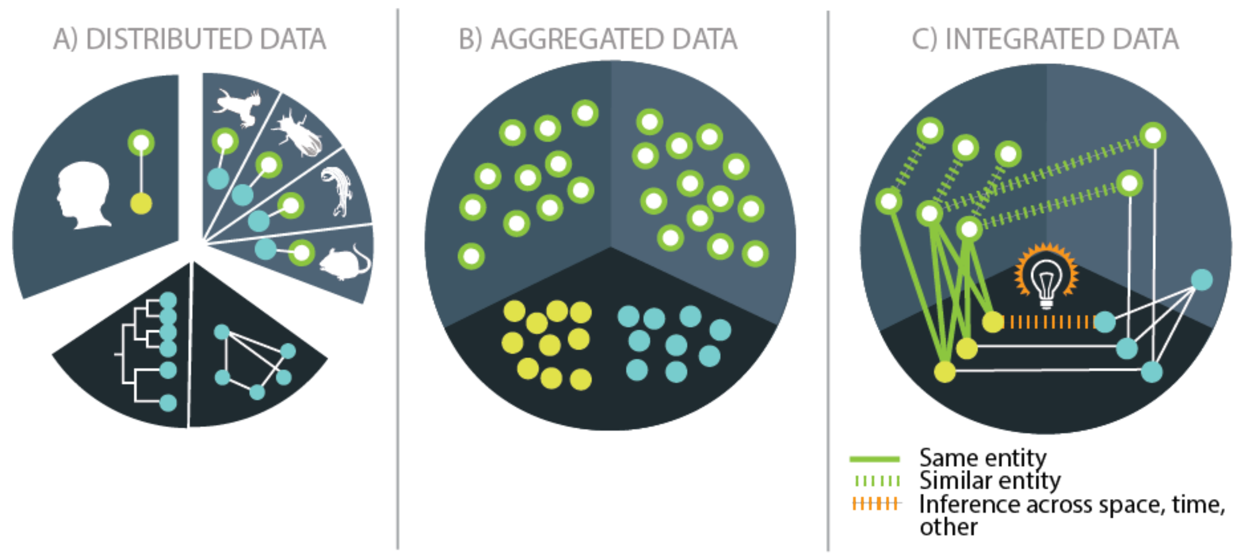

Figure 4: Reintegrating biology is about more than data aggregation. A) Data are distributed across sub-disciplines using bespoke standards. We learn different things from different data sources, but an entire class of big-picture insights are out of reach. B) Data and metadata standards allow us to aggregate data within a sub-discipline, leading to further insight and improved data discoverability, but full integration is still elusive. C) Semantic integration (e.g., ontologies, See Box 1) can provide the missing links across sub-disciplines providing an integrated view that advances our understanding. Figure adapted from [103].

Several initiatives have been developed to promote interoperable, trustworthy data repositories. For example CoreTrustSeal promotes sustainable and trustworthy data repositories by supporting their evaluation [104]; while GO-FAIR is a community-development-focused initiative to aid investigators in making their data FAIR [105]. In addition to these high-level community awareness efforts, a foundational standard is a concrete way to increase interoperability and integration across repositories. One example in biology is the Biolink Model, a graph-oriented data model that was developed for a biomedical use case, but is being extended to the rest of biology [106]. The advantage of such a foundational standard for repositories is that the basic elements (also known as semantic types) of biology (such as a gene or organism) and the relationships between them are identified consistently across resources. This makes integration across repositories much easier (Fig 5). A foundational standard is not meant to replace an existing repository data model; it is meant to provide defined semantic types for annotating a repository data model to aid automated repository integration. Foundational standards, such as the Biolink Model, add machine actionable links across repositories in addition to policies moving toward an integrated, global network of repositories and the data within them.

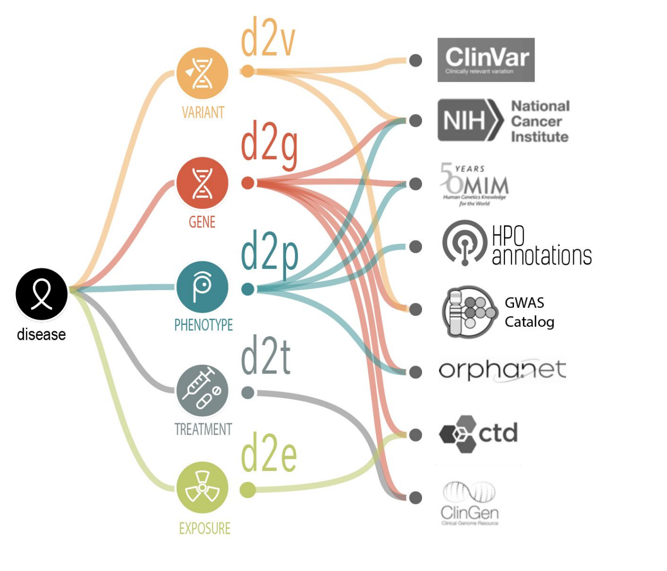

Figure 5: Foundational standards facilitate data integration across biological data repositories. Data repositories are often restricted to a specific sub-discipline or data type. This means that in order to get a complete picture of a problem, data must be integrated across repositories. A foundational data model, such as the Biolink Model, can facilitate integration by providing semantic types for basic biological elements, such as disease, phenotype, gene, etc [106].

Different disciplines are at different stages of readiness for data-centered collaboration. A foundational step is the development of domain-specific standards that describe data and metadata, documenting formats, content, protocols, and vocabularies. There are many different types of standards and they have been described in great detail elsewhere [41]. Progress will require development of standards using a community-driven, consensus-building approach, rather than a top-down approach, that allows each community of practice to develop their own standards [107]. For example, the Genome Standards Consortium (GSC) has established minimum reporting standards for sequence information for the genomics community [108]. Once a community has developed the first version of a standard for representing data and metadata the next need is a standard for integrating data, which can include data exchange standards (e.g., Darwin Core [109] and GA4GH Phenopackets [110]) and domain-specific ontologies (Box 1).

Box 1: What is an ontology?
An ontology is a classification of concepts in a field of knowledge, or a domain, such as organisms or anatomical entities. Concepts are hierarchically arranged, and formally defined

in a human readable format (using text definitions) and computer/machine readable format (encoded with a knowledge representation language like RDFS, OWL, OBO.) In addition, the relationships between concepts are defined, which allows disparate data types to be connected in a formal way. For example, a neuron is a type of brain cell and a brain cell is part of the brain, where neuron, brain cell and brain are concepts and 'is a' and 'part of' are relationships. Ontologies are often used to organize heterogeneous data because their hierarchical organization can overcome problems of scale, granularity, and term heterogeneity. For example, by including data annotated with the terms ‘hippocampus’ and ‘cerebellum’ in a search for data using the keyword ‘brain’. Ontologies form the basis of semantic technology which also include software for creating new links between formerly disparate data sets and integrating data that are not perfectly aligned. These approaches make it possible to overcome some of the integration challenges described above.

Successful standards require a sustained, iterative process of continued development that allows for changes to the standard and updates to the data they describe. This process is typically codified within a governance document that describes the process for updating the standard, resolving disputes, document management, versioning, etc. Examples of successful standards governing bodies include TDWG [111], GSC [112], ESIP [113], OGC [114], INCF [115], CIDOC CRM [116], and H7 FHIR [117]. Effective governance, including a Code of Conduct, can make a big difference in whether or not members feel welcome and that their work is 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.

When investments in standards development have been made, innovations and insights have followed [118,119]; however, such integrative efforts are not the norm in biology. Many subdisciplines have no community standards and the efforts that are taking place are often parochial in nature. For example, despite the existence of ecology data standards, the majority of data in repositories serving ecology and evolution do not have sufficient metadata for reuse [120]. As a result, the critical next steps will be different for different sub-disciplines depending on their readiness for data-collaboration. *We recommend creating, where absent, and supporting, where existing, organizations for the bottom-up development of standards and their governance, using or augmenting existing standards where possible.* Funding agencies can contribute by requiring projects that generate or disseminate data to include support for contribution and adherence to standards. The call for bottom-up standards development is not unique [121,122]. Every sub-discipline community will have to assess where they are in their data collaboration readiness level in order to know the most effective first steps, i.e., forming a standards body, developing an ontology, etc.

<u>Large-scale empirical data collection and monitoring coordination</u>. One way to address integration barriers due to differences in collection methodology and coverage is to stage global efforts to collect the same data using standard protocols, such as NEON [123], IOOS [124], Ocean Sampling Day [125], PhenX [126], and the Census of Marine Life [127]. In addition to

gathering large amounts of homogeneous data, these projects can compare data collected using purposefully different protocols and analyzed by different labs. This exposes variability and results in strategies to mitigate variability. *We recommend identifying opportunities where large-scale data collection and monitoring programs have the greatest potential for reintegrating biology and building on these opportunities.* In some cases, this may include identifying existing monitoring or standardization efforts on which to build.

Automated data curation and integration. With the advent of high-throughput tools for measuring a multitude of biological system aspects, the amount of data is growing exponentially, but data integration and curation tools have not increased in parallel. Tool development is needed, with a particular want in automated solutions to accelerate the work of curation and integrating data across types and scales [13,128]. For example, in order to integrate data sets with non-standard metadata, extra work is needed to identify and relate data types consistently. Machine learning (ML) and artificial intelligence (AI) can help to fill data gaps and create metadata [129]. *We recommend the development and refinement of algorithms for automated metadata creation and format conversion with documentation, provenance, and user interfaces for human-mediated quality assurance.* High-priority automated tasks include named entity recognition, data and semantic typing, protocol detection, and transformation across formats, methods, and units. Some algorithms of this type already exist [130,131], but have not received wide adoption because of problems with usability, sustainability, and discoverability of the tools.

Full transparency of curated data quality. Essential to data integration is user trust in the quality, completeness, and fitness for purpose of the data to be integrated. Repositories, including museums and libraries, are traditional stewards of this trust [132,133]. Trust is complex and sensitive and implies much more than whether the data are right or wrong, but is a consequence of the interactions between data providers, aggregators, users, and repositories [134]. Transparency in all operations is key. A repository that is transparent about process and errors is more trustworthy than a repository that is opaque on these matters. Users must trust that a repository is capable of preserving data and will be persistent, which implies transparency about strategic and business planning and budgeting. Users must have access to provenance information and thorough metadata in order to assess quality and fitness for use [9,44,48,135,136], often using visualizations and studying the workflow used to generate the data [9,44]. *We recommend a policy of full transparency for repositories, aggregators, and integrators that includes documentation of all methods, provenance, processing, modeling, and formats and a reproducible pathway through which users contribute back to data sources [85,133,137,138]*. Such a policy would exceed current industry standards [32]. In this model, all players are considered equal partners in data stewardship and reuse. Essential for achieving this are community standards (discussed above) and a robust system of versioning, provenance, and identifiers [139,140]. Similarly, associating contributor roles to every step is possible; the Contributor Attribution Model and associated Contribution Role Ontology aim to support a greater documentation and understanding of all contributors in the provenance chain [141,142]. Micro-annotation or nanopublication, wherein metadata are associated with individual data atoms (smallest usable elements), can further underpin a system of provenance and attribution tracking where credit cascades through the long pathway of content flow [143,144].

Efforts such as FilteredPush, a tool for distributed data annotation, [145,146] are foundational infrastructure for increasing transparency in data stewards.

Biological data managers and curators. Making full use of rapidly changing technology demands a level of expertise that is difficult to obtain in addition to in-depth biological knowledge. As a result, research teams will increasingly need to include members whose specialty is data management [147]. A professional development structure needs to be established for data curators and information scientists who specialize in data stewardship in order to nurture their careers as part of a research team [128]. *We recommend the development of a much more robust, academic career path for data professionals and recognition of the value of standards and data sharing activities in the context of existing career paths.* The professionalization of data stewardship within academic science will develop incentives for sharing and collaboration more broadly.

Long-term support for data. The current strategy for funding scientific research leaves most data unsupported after the life of the project which generated it. It is an essential but often overlooked aspect of data integration that data are preserved over the long-term. Repositories, including museums and libraries, have the knowledge and expertise to provide sustainable preservation of data after the life of a project [20], but many data repositories accommodate only a single sub-discipline or data type. As a result, many data sets do not have a path to long-term preservation in a trustworthy, curated data repository [32] and researchers are not happy with current resources for long-term data support [75]. Without a repository, researchers have to store their own data locally and the probability that one of these data sets is available decreases by 17% per year [148]. This was due primarily to data sets being held by originators whose contact information could not be found, kept on inaccessible storage media, or just being lost [148]. The lack of preservation resources for many types of research data places much of our collective knowledge in jeopardy, wastes the time and funding that was used to create that knowledge, delays the future insights that might be drawn from that data, and decreases our ability to engage in reproducible science.

Designing, launching, and sustaining a repository is non-trivial, but best practices for administrative planning, data curation, and evolving the infrastructure to meet changing user needs have been published [149–151]. Keys to success include a sound business model, adding value with tools and services, and remaining relevant by responding to user needs [87,152]. An important way for a repository to build value is through integration with other data sets and types [9,15,46,89,153], facilitating curation in a global context [128], and providing mechanisms for updating data sets as standards change over time [154]. For example, biodiversity data must be updated as taxonomies change, but strategies for automated updates and sharing new alignments are an active area of research [155,156]. The resulting, integrated, high-value data set will meet cutting-edge research needs and, as a result, provide motivation for continued support. *We recommend the development of reproducible workflows that support data and metadata from creation to preservation in an accessible repository that is part of a dynamic, global infrastructure of tools and services for data curation and provenance.* A line of funding separate to the discovery-focused research paradigm will be required to assemble and sustain the infrastructure to provide this path for all data. Advocacy by and involvement of international organizations like the International Science Council could help secure this funding.

Call To Action To Reintegrate Biology

Reintegrating biology will require contributions from international science organizations, funding agencies, researchers, and universities that go beyond making data open. With data integration, new opportunities for discovery will be created, tools will emerge to address problems with greater scale and scope, and results will be more reproducible and of improved quality [157]. The data themselves become a resource that can be licensed to support a Data as a Service (DaaS) model [158] with a combination of social and technical infrastructure innovations for frictionless reuse. The biggest hurdle is motivating the sustained community participation required to develop and implement data integration solutions, but this is achievable, as evidenced by the disciplines that have successfully developed and maintained active, community-driven standards. The key to success is recognition (by a critical mass of community members) of a persistent research hurdle that can only be surpassed by large-scale, repeatable data integration. The majority of researchers agree that lack of access to data generated in other laboratories is an impediment to scientific progress [89], and data sharing practices have become more widespread over the past decade [3], suggesting that this recognition has started. More emphasis can be placed on the tangible benefits of reusing data over re-collecting data and quantitative models have been developed for calculating time efficiencies and financial savings [82,159]. It is important to shift opinions about data so that they are viewed as a valuable commodity rather than a costly burden.

Below we list some concrete actions individual researchers can take that build on the decades of effort to improve data sharing and reuse to support the global reintegration of biology. The solutions proposed here will not immediately eliminate the challenges, but will start a journey of incremental, community-driven progress that will require periodic reassessment to ensure the vision and goals are still valuable. There is an urgent need to develop a plan for reintegrating biological data that can address scientific goals, but also addresses the social, political, and infrastructural issues that impede progress.

Eight things researchers can do to reintegrate biology:

- **Design for reuse.** In planning all data generation or analyses, plan *a priori* for reusability (e.g. identifiers, provenance, persistence, etc.)
- **Push data.** Move all data onto a path to a trusted and sustainable data repository whenever possible.
- **Declare licensing.** Choose the most open license possible and ensure that legal reuse of the data is clearly indicated.
- **Use standards.** Identify existing or help develop new standards and apply them to your data.
- **Cite data.** Participate in and utilize data citation guidelines and metrics that transcend discipline, similar to publication citation and metrics.
- **Attribute.** Reintegrating biology will require many hands with different types of expertise; attribute all contributors where possible. This includes tracking contributions to integrated data sets and other integrative artifacts.

- **Create a culture of sharing.** Demonstrate the value of collaborative, interdisciplinary, and integrative practices for students and colleagues in search committees, promotion guidelines, in professional societies, on your own CV, and in citation and attribution.
- **Elevate knowledge work.** Create awareness of software development and knowledge work (such as data modeling, curation, quality assurance) in existing fields and as a new professional domain. Participate in or initiate community efforts to support enhanced training, participation in, attribution of, and metrics for such activities and artifacts.

Summary

The reintegration of the sub-disciplines of biology, and the accompanying insights into the rules of life, require the integration of data across diverse types. Without good data management practices and data science, integration at scale becomes nearly intractable and puts solutions to societal problems out of reach. Significant investment is required to develop data standards, best practices, new mathematical approaches, and a shift in professional incentives that can assist in overcoming the barriers to data integration. Funding agencies can help by specifically supporting - and integrating - efforts to create community-driven data standards and interdisciplinary data architectures. Universities and Institutions can help by rewarding non-traditional activities such as data sharing, interdisciplinary integration, standards development, curation, and other knowledge work. These investments will see a return in the form of increased usability, impact, and marketability of data through a DaaS model. Integration has been focused on human-centric strategies aimed at expanding researcher networks. We need to invest just as much effort into data-centric strategies that expand networks of interoperable data. Addressing these challenges will form a solid observational basis to answer current big questions in biology and contribute science-based solutions to the most pressing social and environmental problems.

Acknowledgements

The authors would like to thank the National Science Foundation for creating the collaborative space to draft this manuscript and our fellow BII workshop participants for feedback and insights. The authors would also like to thank Nicole Vasilevksy 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. Franklin J, Serra-Diaz JM, Syphard AD, Regan HM. Big data for forecasting the impacts of global change on plant communities. Glob Ecol Biogeogr. 2017;26: 6–17.

3. Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, et al. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. PLoS One. 2015;10: e0134826.

4. Edwards PN, Mayernik MS, Batcheller AL, Bowker GC, Borgman CL. Science friction: data,

metadata, and collaboration. Soc Stud Sci. 2011;41: 667–690.

5. McKiernan EC, Bourne PE, Brown CT, Buck S, Kenall A, Lin J, et al. How open science helps researchers succeed. Elife. 2016;5: e16800.

6. Piwowar HA, Day RS, Fridsma DB. Sharing detailed research data is associated with increased citation rate. PLoS One. 2007;2: e308.

7. Nichols TE, Das S, Eickhoff SB, Evans AC, Glatard T, Hanke M, et al. Best practices in data analysis and sharing in neuroimaging using MRI. Nat Neurosci. 2017;20: 299–303.

8. Cranston K, Harmon LJ, O'Leary MA, Lisle C. Best practices for data sharing in phylogenetic research. PLoS Curr. 2014;6. doi:10.1371/currents.tol.bf01eff4a6b60ca4825c69293dc59645

9. 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 -. Ecol Inform. 2012;11: 25–33.

10. Thessen AE, Patterson DJ. Data issues in the life sciences. Zookeys. 2011;150. doi:10.3897/zookeys.150.1766

11. Pronk TE, Wiersma PH, Weerden A van, Schieving F. A game theoretic analysis of research data sharing. PeerJ. 2015;3: e1242.

12. Poole AH. How has your science data grown? Digital curation and the human factor: a critical literature review. Archival Science. 2015;15: 101–139.

13. 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.

14. Fecher B, Friesike S, Hebing M. What drives academic data sharing? PLoS One. 2015;10: e0118053.

15. Wallis JC, Rolando E, Borgman CL. If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PLoS One. 2013;8: e67332.

16. Chawinga WD, Zinn S. Global perspectives of research data sharing: A systematic literature review. Libr Inf Sci Res. 2019;41: 109–122.

17. Bertagnolli MM, Sartor O, Chabner BA, Rothenberg ML, Khozin S, Hugh-Jones C, et al. Advantages of a Truly Open-Access Data-Sharing Model. N Engl J Med. 2017;376: 1178–1181.

18. 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.

19. Faniel IM, Zimmerman A. Beyond the Data Deluge: A Research Agenda for Large-Scale Data Sharing and Reuse. International Journal of Digital Curation. 2011;6: 58–69.

20. 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.

21. Thessen AE. Adoption of machine learning techniques in Ecology and Earth Science. One Ecosystem. 2016;1: e8621.

22. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nat Methods. 2018;15: 233–234.

23. Hardisty A, Roberts D, The Biodiversity Informatics Community. A decadal view of biodiversity informatics: challenges and priorities. BMC Ecol. 2013;13. doi:10.1186/1472-6785-13-16

24. Thessen AE, Fertig B, Jarvis JC, Rhodes AC. Data Infrastructures for Estuarine and Coastal Ecological Syntheses. Estuaries Coasts. 2016;39: 295–310.

25. 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.

26. 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.

27. HealthITAnalytics. AMIA: Health Informatics Can Help Overcome the Big Data “Deluge.” In: HealthITAnalytics [Internet]. 13 Jul 2017 [cited 26 May 2020]. Available: https://healthitanalytics.com/news/amia-health-informatics-can-help-overcome-the-big-data-deluge

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

29. Stroud JT, Bush MR, Ladd MC, Nowicki RJ, Shantz AA, Sweatman J. Is a community still a community? Reviewing definitions of key terms in community ecology. Ecol Evol. 2015;5: 4757–4765.

30. 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.

31. Leonelli S. The challenges of big data biology. Elife. 2019;8. doi:10.7554/eLife.47381

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

33. Carmona CP, de Bello F, Mason NWH, Lepš J. Traits Without Borders: Integrating Functional Diversity Across Scales. Trends Ecol Evol. 2016;31: 382–394.

34. Ryan PJ, McKenzie NJ, O’Connell D, Loughhead AN, Leppert PM, Jacquier D, et al. Integrating forest soils information across scales: spatial prediction of soil properties under

Australian forests. For Ecol Manage. 2000;138: 139–157.

35. Webb TJ, Vanden Berghe E, O'Dor R. Biodiversity's big wet secret: the global distribution of marine biological records reveals chronic under-exploration of the deep pelagic ocean. PLoS One. 2010;5: e10223.

36. Science Results | SDSS. [cited 29 May 2020]. Available: https://www.sdss.org/science/

37. Blazquez D, Domenech J. Big Data sources and methods for social and economic analyses. Technol Forecast Soc Change. 2018;130: 99–113.

38. National Centers for Environmental Information (NCEI). [cited 29 May 2020]. Available: https://www.ncei.noaa.gov/

39. Gewin V. Data sharing: An open mind on open data. Nature. 2016;529: 117–119.

40. Data sharing and the future of science. Nat Commun. 2018;9: 2817.

41. Standards. [cited 26 May 2020]. Available: http://rd-alliance.github.io/metadata-directory/standards/

42. 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.

43. 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;15: e2001414.

44. 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.

45. Gross T, Taylor AG, Joudrey DN. Still a Lot to Lose: The Role of Controlled Vocabulary in Keyword Searching. Cataloging & Classification Quarterly. 2015;53: 1–39.

46. Perspectives K. Data dimensions: disciplinary differences in research data sharing, reuse and long term viability. SCARP Synthesis Study, Digital Curation Centre. 2010.

47. Gregory KM, Cousijn H, Groth P, Scharnhorst A, Wyatt S. Understanding data search as a socio-technical practice. J Inf Sci Eng. 2019; 0165551519837182.

48. Gregory K, Groth P, Cousijn H, Scharnhorst A, Wyatt S. Searching Data: A Review of Observational Data Retrieval Practices in Selected Disciplines. J Assoc Inf Sci Technol. 2019;70: 419–432.

49. Thessen AE, Patterson DJ, Murray SA. The Taxonomic Significance of Species That Have Only Been Observed Once: The Genus Gymnodinium (Dinoflagellata) as an Example. PLoS One. 2012;7. doi:10.1371/journal.pone.0044015

50. Xue Y, Bogdan P. Constructing Compact Causal Mathematical Models for Complex Dynamics. Proceedings of the 8th International Conference on Cyber-Physical Systems.

New York, NY, USA: ACM; 2017. pp. 97–107.

51. Ghorbani M, Jonckheere EA, Bogdan P. Gene Expression Is Not Random: Scaling, Long-Range Cross-Dependence, and Fractal Characteristics of Gene Regulatory Networks. Front Physiol. 2018;9: 1446.

52. Peng CK, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL. Mosaic organization of DNA nucleotides. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics. 1994;49: 1685–1689.

53. Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, Matsa ME, Peng CK, et al. Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics. 1995;51: 5084–5091.

54. Ivanov PC, Amaral LA, Goldberger AL, Havlin S, Rosenblum MG, Struzik ZR, et al. Multifractality in human heartbeat dynamics. Nature. 1999;399: 461–465.

55. Goldberger AL, Amaral LAN, Hausdorff JM, Ivanov PC, Peng C-K, Stanley HE. Fractal dynamics in physiology: alterations with disease and aging. Proc Natl Acad Sci U S A. 2002;99 Suppl 1: 2466–2472.

56. Liu Y, Schwalm CR, Samuels-Crow KE, Ogle K. Ecological memory of daily carbon exchange across the globe and its importance in drylands. Ecol Lett. 2019;22: 1806–1816.

57. França LGS, Souza França LG, Vivas Miranda JG, Leite M, Sharma NK, Walker MC, et al. Fractal and Multifractal Properties of Electrographic Recordings of Human Brain Activity: Toward Its Use as a Signal Feature for Machine Learning in Clinical Applications. Frontiers in Physiology. 2018. doi:10.3389/fphys.2018.01767

58. Racz FS, Stylianou O, Mukli P, Eke A. Multifractal Dynamic Functional Connectivity in the Resting-State Brain. Front Physiol. 2018;9: 1704.

59. Bogdan P, Deasy BM, Gharaibeh B, Roehrs T, Marculescu R. Heterogeneous structure of stem cells dynamics: statistical models and quantitative predictions. Sci Rep. 2014;4: 4826.

60. Bassingthwaighte JB, Liebovitch LS, West BJ. Fractal Physiology. Springer; 2013.

61. Nonnenmacher TF, Losa GA, Weibel ER. Fractals in Biology and Medicine. Birkhäuser; 2013.

62. Xue Y, Rodriguez S, Bogdan P. A spatio-temporal fractal model for a CPS approach to brain-machine-body interfaces. 2016 Design, Automation Test in Europe Conference Exhibition (DATE). 2016. pp. 642–647.

63. Chen J, Xiao Y, Gai Z, Li R, Zhu Z, Bai C, et al. Reproductive toxicity of low level bisphenol A exposures in a two-generation zebrafish assay: Evidence of male-specific effects. Aquat Toxicol. 2015;169: 204–214.

64. Ghorbani M, Bogdan P. A cyber-physical system approach to artificial pancreas design. 2013 International Conference on Hardware/Software Codesign and System Synthesis

(CODES+ISSS). 2013. pp. 1–10.

65. Albert CH, Grassein F, Schurr FM, Vieilledent G, Violle C. When and how should intraspecific variability be considered in trait-based plant ecology? Perspect Plant Ecol Evol Syst. 2011;13: 217–225.

66. Geiler-Samerotte KA, Bauer CR, Li S, Ziv N, Gresham D, Siegal ML. The details in the distributions: why and how to study phenotypic variability. Curr Opin Biotechnol. 2013;24: 752–759.

67. Lianou A, Koutsoumanis KP. Strain variability of the behavior of foodborne bacterial pathogens: a review. Int J Food Microbiol. 2013;167: 310–321.

68. Nurden AT, Fiore M, Nurden P, Pillois X. Glanzmann thrombasthenia: a review of ITGA2B and ITGB3 defects with emphasis on variants, phenotypic variability, and mouse models. Blood. 2011;118: 5996–6005.

69. Pfennig DW, Wund MA, Snell-Rood EC, Cruickshank T, Schlichting CD, Moczek AP. Phenotypic plasticity's impacts on diversification and speciation. Trends Ecol Evol. 2010;25: 459–467.

70. Patterson DJ, Cooper J, Kirk PM, Pyle RL, Remsen DP. Names are key to the big new biology. Trends Ecol Evol. 2010;25: 686–691.

71. Patterson D, Mozzherin D, Shorthouse DP, Thessen A. Challenges with using names to link digital biodiversity information. Biodiversity Data Journal. 2016;4. doi:10.3897/BDJ.4.e8080

72. Berendsohn WG, Geoffroy M. Networking. Taxonomic. Concepts.—. Uniting. without."Unitary-ism." Biodiversity Databases. CRC Press; 2016. pp. 25–34.

73. Perrier L, Blondal E, MacDonald H. The views, perspectives, and experiences of academic researchers with data sharing and reuse: A meta-synthesis. PLoS One. 2020;15: e0229182.

74. Bryan Heidorn P. Shedding Light on the Dark Data in the Long Tail of Science. Libr Trends. 2008;57: 280–299.

75. 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.

76. The Research Parasite Awards. [cited 15 Dec 2019]. Available: https://researchparasite.com/

77. Longo DL, Drazen JM. Data Sharing. The New England journal of medicine. 2016. pp. 276–277.

78. DataCite. Welcome to DataCite. 2018. Available: https://datacite.org/

79. Penev L, Mietchen D, Chavan V, Hagedorn G, Remsen D, Smith V, et al. Pensoft data publishing policies and guidelines for biodiversity data. Pensoft Publ. 2011. Available:

https://www.researchgate.net/profile/Lyubomir_Penev/publication/265422943_Pensoft_Data_Publishing_Policies_and_Guidelines_for_Biodiversity_Data/links/5410c8a00cf2f2b29a411603/Pensoft-Data-Publishing-Policies-and-Guidelines-for-Biodiversity-Data.pdf

80. 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. doi:10.29173/iq240

81. 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.

82. Vines TH, Andrew RL, Bock DG, Franklin MT, Gilbert KJ, Kane NC, et al. Mandated data archiving greatly improves access to research data. FASEB J. 2013;27: 1304–1308.

83. Froese R, Lloris D, Opitz S. Scientific data in the public domain. ACP-EU Fish Res Rep. 2003;14: 267–271.

84. Henty M, Weaver B, Bradbury S, Porter S, Others. Investigating data management practices in Australian universities. 2008. Available: https://openresearch-repository.anu.edu.au/handle/1885/47627

85. Costello MJ. Motivating Online Publication of Data. Bioscience. 2009;59: 418–427.

86. Savage CJ, Vickers AJ. Empirical study of data sharing by authors publishing in PLoS journals. PLoS One. 2009;4: e7078.

87. Smith VS. Data publication: towards a database of everything. BMC Res Notes. 2009;2: 113.

88. Cragin MH, Palmer CL, Carlson JR, Witt M. Data sharing, small science and institutional repositories. Philos Trans R Soc Lond A. 2010;368: 4023–4038.

89. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, et al. Data sharing by scientists: practices and perceptions. PLoS One. 2011;6: e21101.

90. Rebecca J. Griffiths, Nancy L. Maron, Kevin M. Guthrie. Sustainability and Revenue Models for Online Academic Resources. 2008.

91. Blair J, Gwiazdowski R, Borrelli A, Hotchkiss M, Park C, Perrett G, et al. Towards a catalogue of biodiversity databases: An ontological case study. BDJ. 2020;8: e32765.

92. Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Sustainable Economics for a Digital Planet: Ensuring Long-term Access to Digital Information. 2010. Available: http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

93. Sayogo DS, Pardo TA. Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. Gov Inf Q. 2013;30: S19–S31.

94. The Carpentries. In: The Carpentries [Internet]. [cited 15 Dec 2019]. Available:

https://carpentries.org/

95. Oxenham S. Legal maze threatens to slow data science. Nature. 2016;536: 16–17.

96. Carbon S, Champieux R, McMurry J, Winfree L, Wyat LR, Haendel M. A Measure of Open Data: A Metric and Analysis of Reusable Data Practices in Biomedical Data Resources. bioRxiv. 2018. p. 282830. doi:10.1101/282830

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

98. 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.

99. 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.

100. (Re)usable Data Project. [cited 29 May 2020]. Available: https://reusabledata.org/

101. 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.

102. Haendel M, Su A, McMurry J. FAIR-TLC: Metrics to Assess Value of Biomedical Digital Repositories: Response to RFI NOT-OD-16-133. 2016. doi:10.5281/zenodo.203295

103. 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.

104. Corrado EM. Repositories, Trust, and the CoreTrustSeal. Technical Services Quarterly. 2019;36: 61–72.

105. GO FAIR. In: GO FAIR [Internet]. [cited 20 May 2020]. Available: https://www.go-fair.org

106. biolink-model. Github; Available: https://github.com/biolink/biolink-model

107. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25: 1251–1255.

108. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol. 2011;29: 415–420.

109. Wieczorek J, Bloom D, Guralnick R, Blum S, Doring M, Giovanni R, et al. Darwin Core: An evolving community-developed biodiversity data standard. PLoS One. 2012;7: e29715.

110. Welcome to the documentation for phenopackets-schema! — phenopackets-schema

10.0.0 documentation. [cited 31 May 2020]. Available: https://phenopackets-schema.readthedocs.io/en/latest/

111. Biodiversity Information Standards (TDWG). Biodiversity Information Standards (TDWG). [cited 18 Dec 2019]. Available: https://www.tdwg.org/

112. Field D, Sterk P, Kottmann R, De Smet JW, Amaral-Zettler L, Cochrane G, et al. Genomic standards consortium projects. Stand Genomic Sci. 2014;9: 599–601.

113. ESIP | Connecting Science, Data and Users. [cited 18 Dec 2019]. Available: https://www.esipfed.org/

114. Welcome to The Open Geospatial Consortium | OGC. [cited 18 Dec 2019]. Available: https://www.opengeospatial.org/

115. Standards and Best Practices organisation for open and FAIR neuroscience | INCF - International Neuroinformatics Coordinating Facility. [cited 18 Dec 2019]. Available: https://www.incf.org/

116. Home | CIDOC CRM. [cited 20 Mar 2020]. Available: http://www.cidoc-crm.org/

117. Overview - FHIR v4.0.1. [cited 15 Dec 2019]. Available: https://www.hl7.org/fhir/overview.html

118. Mungall CJ, McMurry JA, Köhler S, Balhoff JP, Borromeo C, Brush M, et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res. 2017;45: D712–D722.

119. Haendel MA, Vasilevsky N, Brush M, Hochheiser HS, Jacobsen J, Oellrich A, et al. Disease insights through cross-species phenotype comparisons. Mamm Genome. 2015;26: 548–555.

120. Roche DG, Kruuk L, Lanfear R, Binning SA. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? PLoS Biol. 2015;13: e1002295.

121. 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.

122. Aubin I, Cardou F, Boisvert-Marsh L, Garnier E, Strukelj M, Munson AD. Managing data locally to answer questions globally: The role of collaborative science in ecology. Bello F, editor. J Veg Sci. 2020;31: 509–517.

123. Schimel D, Hargrove W, Hoffman F, MacMahon J. NEON: a hierarchically designed national ecological network. Front Ecol Environ. 2007;5: 59–59.

124. Brown V. Technologies converge to make integrated ocean observing system a reality. Environ Sci Technol. 2004;38: 198A–199A.

125. Ocean Sampling Day | Micro B^3. [cited 19 Dec 2019]. Available: https://www.microb3.eu/osd.html

126. Hamilton CM, Strader LC, Pratt JG, Maiese D, Hendershot T, Kwok RK, et al. The PhenX Toolkit: get the most from your measures. Am J Epidemiol. 2011;174: 253–260.

127. Ausubel JH, Trew C, Waggoner PE, Others. First Census of Marine Life 2010: Highlights of a decade of discovery. First census of marine life 2010: highlights of a decade of discovery. 2010. Available: https://www.cabdirect.org/cabdirect/abstract/20113162027

128. 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.

129. Bionetworks S. Synapse | Sage Bionetworks. [cited 31 May 2020]. Available: https://www.synapse.org/

130. Mozzherin D, Myltsev AA, Patterson D. Finding scientific names in Biodiversity Heritage Library, or how to shrink Big Data. BISS. 2019;3: e35353.

131. Mozzherin DY, Myltsev AA, Patterson DJ. “gnparser”: a powerful parser for scientific names based on Parsing Expression Grammar. BMC Bioinformatics. 2017;18: 279.

132. Quality assurance and Intellectual Property Rights in advancing biodiversity data publication. [cited 25 Mar 2020]. Available: http://www.gbif.org/resource/80818

133. 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.

134. Franz NM, Sterner BW. To increase trust, change the social design behind aggregated biodiversity data. Database . 2018;2018. doi:10.1093/database/bax100

135. Weber NM, Baker KS, Thomer AK, Chao TC, Palmer CL. Value and context in data use: Domain analysis revisited. Proc Am Soc Info Sci Tech. 2012;49: 1–10.

136. Bishop BW, Hank C, Webster J, Howard R. Scientists’ data discovery and reuse behavior: (Meta)data fitness for use and the FAIR data principles. Proceedings of the Association for Information Science and Technology. 2019;56: 21–31.

137. Faith D, Collen B, Ariño A, Koleff PKP, Guinotte J, Kerr J, et al. Bridging the biodiversity data gaps: Recommendations to meet users’ data needs. Biodivers Inf. 2013;8. Available: https://www.jcel-pub.org/index.php/jbi/article/view/4126

138. Soranno PA, Bissell EG, Cheruvelil KS, Christel ST, Collins SM, Fergus CE, et al. Building a multi-scaled geospatial temporal ecology database from disparate data sources: fostering open science and data reuse. Gigascience. 2015;4: 28.

139. Guralnick RPP, Cellinese N, Deck J, Pyle RLL, Kunze J, Penev L, et al. Community Next Steps for Making Globally Unique Identifiers Work for Biocollections Data. Zookeys. 2015;494: 133–154.

140. Gil Y, David CH, Demir I, Essawy BT, Fulweiler RW, Goodall JL, et al. Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance. Life Support Biosph Sci. 2017; 388–415.

141. Transitive Credit and JSON-LD. Journal of Open Research Software. 2015;3: 14.

142. Welcome to the Contributor Attribution Model — Contributor Attribution Model documentation. [cited 31 May 2020]. Available: https://contributor-attribution-model.readthedocs.io/en/latest/

143. 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.

144. 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.

145. Morris PJ, Macklin JA, Hanken J, Kelly M, Koehler S, Lowery D, et al. Improving Natural Science Collections data through quality control for research using Kepler workflows embedded in a FilteredPush network. SPNHC 2013. 2013; 33.

146. Morris PJ, Kelly MA, Lowery DB. Filtered Push: annotating distributed data for quality control and fitness for use analysis. AGU Fall Meeting. 2009. Available: http://adsabs.harvard.edu/abs/2009AGUFMIN34B..08M

147. Yakel E, Faniel IM, Maiorana ZJ. Virtuous and vicious circles in the data life-cycle. 2019. Available: http://informationr.net/ir/24-2/paper821.html

148. 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.

149. Wren JD, Bateman A. Databases, data tombs and dust in the wind. Bioinformatics. 2008;24: 2127–2128.

150. Whyte A. A pathway to sustainable research data services: From scoping to sustainability. In: Pryor G, Jones S, Whyte A, editors. Delivering Research Data Management Services: Fundamentals of Good Practice. Facet Publishing; 2013. pp. 59–88.

151. Rieger OY. Sustainability: Scholarly repository as an enterprise. Bul Am Soc Info Sci Tech. 2012;39: 27–31.

152. Borgman CL. Research Data: Who Will Share What, with Whom, When, and Why? 2010. doi:10.2139/ssrn.1714427

153. Van Horn JD, Gazzaniga MS. Why share data? Lessons learned from the fMRIDC. Neuroimage. 2013;82: 677–682.

154. Yenni GM, Christensen EM, Bledsoe EK, Supp SR, Diaz RM, White EP, et al. Developing a modern data workflow for regularly updated data. PLoS Biol. 2019;17: e3000125.

155. Vaidya G, Lepage D, Guralnick R. The tempo and mode of the taxonomic correction process: How taxonomists have corrected and recorrected North American bird species over the last 127 years. PLoS One. 2018;13: e0195736.

156. Franz NM, Pier NM, Reeder DM, Chen M, Yu S, Kianmajd P, et al. Taxonomic Provenance: Two Influential Primate Classifications Logically Aligned. arXiv [q-bio.PE]. 2014. Available: http://arxiv.org/abs/1412.1025

157. Molloy JC. The Open Knowledge Foundation: open data means better science. PLoS Biol. 2011;9: e1001195.

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

159. Pronk TE. The Time Efficiency Gain in Sharing and Reuse of Research Data. Data Science Journal. 2019. Available: https://datascience.codata.org/article/10.5334/dsj-2019-010/

Author Contributions

AET: Conceptualization, Writing – original draft
PB: Conceptualization, Writing – original draft
DJP: Conceptualization, Writing – original draft
TC: Conceptualization, Writing – review & editing
CH: Conceptualization, Writing – review & editing
OdL: Conceptualization, Writing – review & editing
MAH: Conceptualization, Writing – review & editing

Competing Interests Statement

The authors declare no competing interests.