Title:
Reintegration of biology through data compilation and integration across the diverse and distinct fields of life sciences: Challenges, requirements, opportunities and actions

Authors:
Anne Thessen, Cesar Hinojo, Theresa Casey, Paul Bogdan, Orlando de Lange

Introduction

Life on Earth exists as diverse, interacting biological systems to include the only extant form of humans, *Homo sapiens*. Nearly eight billion humans currently inhabit the planet. Feeding these people, preventing the spread of contagious disease, and dealing with the demands of the expanding human population on Earth's resources and biodiversity call for scientists to come together across disciplines to understand potential consequences and develop strategies that foster the health of all living systems.

The National Science Foundation is spearheading efforts to transform our understanding of life and the interdependence of all its forms by calling for the reintegration of biology into a unified and integrated field. Fundamental to integrating the silos of biology is the compilation, mining and integration of data. Data are essential for testing hypotheses, models, and theory as they correspond to the actual observations of the natural phenomena. Integration of data from diverse groups of experts and biological fields has an enormous potential to move forward discovery and understanding of life on Earth. Many of the most pressing societal challenges, such as climate change, biodiversity loss, and sustainable ecosystem management, require the integration of heterogeneous data collected across multiple scales and disciplines. This makes data management, integration, and gap-filling fundamental to the future of scientific advancement.

Data compilation is challenged by the need for development of management tools, services, and mathematical techniques. Moreover, the work of collecting, curating, and analyzing data is still very siloed by discipline. Here, we discuss nine challenges that prevent effective and efficient utilization of biological data across three categories: challenges deriving from the nature of the data, challenges deriving from the nature of biological systems, and challenges deriving from the nature of people (Box 1).

Box 1: Biological Data Integration Challenges

Challenges in the Nature of the Data

- Data are collected on multiple scales: Integrating biological data requires combining data collected at the molecular scale (photosynthesis) with data collected at the global scale (remote sensing). Data about processes that occur in milliseconds need to be combined with processes that occur on geological time.
- Data are highly variable: Data relevant for biology are highly heterogeneous in form, terms used, metadata provided, and protocols used to collect data.
- Data sets have gaps: Life processes and the factors affecting them have been sampled unevenly across the planet and gaps will always be present.

Challenges in the Nature of Biological Systems

- Biological systems are dynamic: The systems that data represent are filled with feedback loops, trajectories, and stochasticity. Data represent a snapshot in time, which makes representing dynamics challenging.

- Biological systems have memory: The state of a system at any point in time is often affected by the conditions of the previous state, to include relative time (minutes, days, years), development and generation. Thus, data from a single instance rarely captures the complete picture of how the system works.

Challenges in the Nature of People

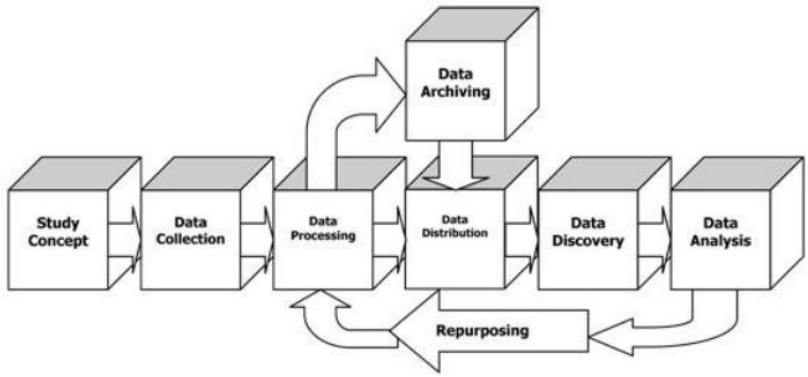
- Discovery: Researchers have trouble finding the data they need for a variety of technological and social reasons including a heavy reliance on word-of-mouth.
- Access: When researchers do find data, they may have trouble accessing it, as sharing platforms are limited and not standardized. Professional incentives offer limited support for data sharing.
- Education: Students and researchers may not have the skills they need to analyse and integrate data of different types and scales.
- Support: Large-scale projects that require a high degree of data integration need public support and funding.

This paper presents a vision for data compilation and integration across the silos of biological disciplines that will enable the pursuit of larger-scale questions still unresolved by previous, more reductionist approaches. Addressing these challenges, even partially, will push biology through a tipping point into new discoveries and make data more accessible to a wider audience. Open access to data and tools has the potential to democratize innovation. Reductionist approaches have led to many exciting discoveries, but now is the time to "zoom out" by connecting diverse data sets to learn and transform our understanding of life on Earth.

Background

Current Landscape

Since the Enlightenment, scientific progress has had a foundation in data and observations. All scientific advancement is based on data gathered from the real world; even models and simulations have to be grounded in reality. Thus, the creation and management of data are essential for progress in scientific knowledge. Data itself has a life cycle. Once gathered, data can be reused and reanalyzed as new discoveries are made. The reanalysis may better explain concepts or phenomena within the new context of understanding of the systems being studied [1–3] (Fig. 1). In order to discover and comprehensively understand more complex scientific systems of greater scope, we will need to be able to integrate larger and more diverse data sets. However, the divergence of biological disciplines and variability of data types are currently inhibiting progress on a class of large-scale, convergent questions about living systems. Researchers have been struggling with the nine challenges outlined in Box 1 for years. Progress is possible now because of recent improvements in computing power, computational methods, and community data standards.



DDI Alliance Structural Reform Group (2004). DDI Combined LIfe Cycle Model. http://www.ddialliance.org/system/files/Concept-Model-WD.pdf

Figure 1: Data Lifecycle. From https://ddialliance.org/sites/default/files/Concept-Model-WD.pdf. Data can be reused and reanalyzed numerous times after its initial collection.

Challenges in the Nature of the Data

Integration of data across disciplinary silos is challenged by the types of data that exist. The entire body of scientific knowledge is based on centuries of observations taken by thousands of researchers for myriad purposes. As a result, data are highly heterogeneous in form, terminology, and protocol. Rich technological sensing and intrinsic multiscale biological dynamics contribute to rich (heterogeneous) data modalities ranging from discrete levels (e.g., gene expressed/not expressed), to continuous values (e.g., protein concentrations), to graph-like structures (e.g., 3D chromatin conformation). Even if the data are very similar in form, different disciplines or even different researchers in the same discipline can use terms ambiguously [4]. Heterogeneity can also arise from the cyclicity and noise of the biological phenomena being measured, which can change randomly or across spatiotemporal scales. Thus, combining data of different types requires special consideration.

Biological systems exhibit multiscale phenomena, and so data are collected over multiple spatial and temporal scales (Fig 2). A mutation in a gene that changes the way a protein functions can affect the behavior of an organism and change the functioning of an ecosystem. It is not always clear how to relate measurements of subcellular structures and molecular processes to measurements of global ecosystem processes. Even measurements of the same process taken at different frequencies can be a challenge to integrate. Moreover, biological systems often behave non-linearly, and so existing data cannot be assumed to represent classic time and space (erdotic) principles. Ignoring these characteristics of a system during data analysis and model building can lead to incorrect predictive models, which impede the definition of and control towards a (good/healthy) performance envelope, [5–7] or obscure tipping points of fragility, disease, and frailty in biological systems.

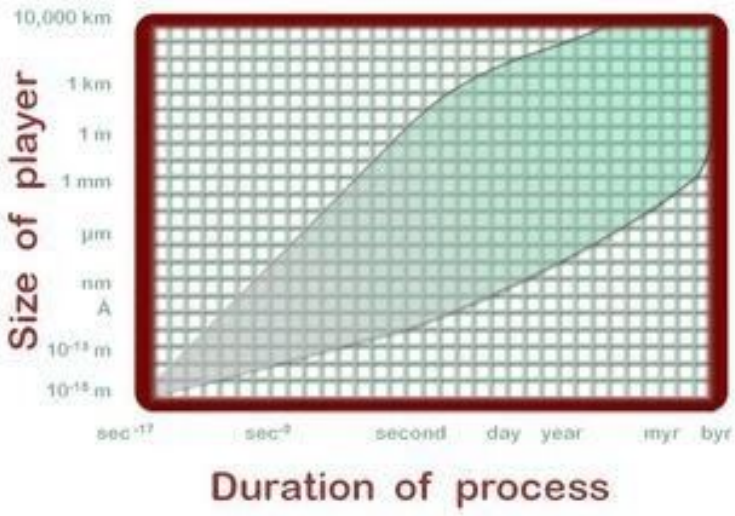


Figure 2: Nature's Envelope. Reused from [8]. Entities (vertical axis) and processes (horizontal axis) occur over a wide variety of scales. The shaded area shows the combination of temporal and spatial scaling in which the dynamics of biological systems occur.

The planet is very unevenly sampled [9]. Uneven sampling is due to limitations in monitoring biological systems. For example, due to limitations in technological sensing, some important variables cannot be measured continuously or even sensed without affecting others. Some of the most important variables are also among the more labor intensive and time consuming to measure. This can result in samples with few replicates that have very low representation of real variation across time, space, and biological levels. There is also difficulty in uniformly sensing biological systems that encompass large spatial dimensions / regions, which can lead to sampling bias or a granularity "mismatch" between data sets. Sampling bias implicit or explicit 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.

Challenges in the Nature of Biological Systems

The systems that data represent are very dynamic and are rarely in steady state equilibrium. They are characterized by multiple concurrent processes that may interact or have cyclicity. Biological systems can change their model structure (i.e., oscillating between various types of linear / nonlinear interdependencies in order to cope with environmental perturbations while ensuring specific biological functionalities) or change the parameters within their model structure over time [10]. Interactions between the different components of a biological system can be complicated many-to-many relationships whose effects can be observed at some scales, but not others. This is reflected, for example, in the gene-to-transcription factor interactions in gene regulatory networks of bacteria [11]. Additionally, systems are monitored in the real world, with uncontrolled (or even unknown) external conditions which affect the data in ways that are not well known. Our ability to understand these systems is limited by the variables we can measure, which in turn is limited by technology and resources. Decisions about what to measure can reflect bias and assumptions rather than biological reality. Often, the true dynamics of these

systems must be inferred from these biased data sets. Dynamic change is a critical component of biological systems that is difficult to capture using current data integration and analysis strategies.

Biological systems exhibit long-range memory and fractal dynamics, meaning that current state depends not only on what is happening now, but also what happened in the past [12–16]. Thus, having all the data from a specific time point is not enough to understand the system. You also need data about what happened the day, week, month, or even year before. Long-term memory is the dynamic outcome of several interactive processes, rather than the result of a linear sequence of events. Accounting for this memory (of organisms and systems) in data, analysis, and interpretation imposes challenges on its own, and thus further challenges the integration of data across disciplines and scales. Consequently, new mathematical and algorithmic techniques need to be developed not only for capturing the mathematical characteristics of biological data (e.g., non-Gaussianity, non-Markovianity), but also to be capable of determining the number of unobserved variables and perturbations [17] in order to construct comprehensive causal predictive models of the rules of life.

Challenges in the Nature of the People

The first steps in data integration is data discovery. The challenges of data discovery begin with the culture of sharing - in biological sciences, academia, and in groups of people in general [18,19]. Whereas meteorology, economics, and astronomy are built on open data, data sharing in the biological sciences is less mature and fraught with social and technical barriers [20–23]. Despite this, 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 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 [20]. In addition, data sharing can democratize access to data types that require expensive equipment, improving access for researchers at small institutions. A robust culture of data sharing has the potential to revolutionize the social aspect of research.

Even if made available, data discovery is still limited both by humans and the computational tools used to find data on the web. Humans are able to identify and interpret a wide variety of contextual cues in order to manage data heterogeneity, but are unable to operate at the scope, scale, and speed needed to cope with the volume of scientific data [24]. Computers are much better able to handle the volume, but cannot cope with heterogeneity or use context as well as a human. Thus, humans increasingly rely on computational agents for data discovery, but without pervasive and consistent use of identifiers [25], data standards [26], metadata standards [27], and controlled vocabularies [28] search tools will not be effective.

Even when students and researchers are able to discover and access all the relevant data, they may not have the skills needed to integrate it correctly or at scale. Proper data integration requires knowledge of the data sets to be integrated (assuming metadata and documentation exists), and frequently requires knowledge of data formats, specific software tools, and computer programming. Informal education, such as the Carpentries [29] (a global community of instructors teaching basic programming and data science skills) has contributed to solving some of the training gaps and online tutorials have filled in others. Universities are beginning to offer data science programs, but these are not always integrated with the traditional subjects. It is still possible to get an education in biological sciences without learning the basics of data management.

Data-driven discovery depends on access and reuse of data at a scale that must reach across projects and laboratories. Investments in data infrastructure such as tools, services and community standards are needed to enable data curation and access, but cannot be accomplished as a minimally funded “broader impacts” aside. This means that projects and funding dedicated to the creation of data infrastructure are needed. Since much of the research and development work in the United States is funded by federal agencies, the support of these endeavors depends on the US Congress and thus the public. Infrastructure to promote data reuse is definitely worth the investment, considering reuse is cheaper than recollection. Even so, these projects can still be a difficult sell for the same reason most infrastructure projects are neglected, i.e., they are not very exciting. An additional problem is the US taxpayers’ attitudes about science, which have been poor (but are improving) and have resulted in several federally-funded projects being targeted by conservative lawmakers [30,31]. The stable and effective cyberinfrastructure necessary for data-driven discovery requires clear demonstration of the return on investment if we are to expect the public to bear the cost. The return on investment is clear, but convincing the US taxpayer in our post-truth era will require communication strategies that are new to many scientists [32].

Proposed Solutions

Solving these problems will require a consistent and dedicated effort to develop new technologies and community research norms. Below, we propose six solutions to address the challenges discussed above.

Community-driven standards for collection, sharing, and analysis. An important barrier to data integration is the vast heterogeneity of the terms, formats, metadata, and protocols. Some scientific communities have made progress by creating community-driven standards using a bottom-up, consensus-building approach (e.g., [33]). For example, the Genome Standards Consortium (GSC) has established minimum reporting standards for sequence information [34] and the Earth Science Information Partnership (ESIP) holds regular meetings to develop standards for data and software [35]. The systematics community has standards for how species are described, how photographs are taken, and how data are shared [36]. Efforts to standardize methods for most commonly measured/used plant traits and techniques in plant

physiological ecology have been made [37,38]. Many of the standards that are used are still very parochial, only used by a few labs or within a very focused sub-discipline.

Biology, with the exception of biomedicine and a few other subdisciplines [33,39–41], has lagged in standards development and could benefit from sustained, coordinated efforts. Where investments in standards development have been made, innovations and insights have followed, e.g., in biomedicine [42,43]. These successful efforts have developed around a standards body with a dedicated mission to develop and maintain standards. We recommend the development of a similar body for the biological science sub-disciplines that do not already have a process for developing and maintaining standards.

Large-scale empirical data collection and monitoring coordination. Two important barriers to data integration are the differences in protocol used to generate the data, and that different datasets may have a different subset of the variables of interest. One way to address this is to stage a global effort to collect data using the same protocols to measure variables of wide interest and importance for biology, such as NEON [44], IOOS [45], Ocean Sampling Day [46], and the Census of Marine Life [47]. In addition to gathering large amounts of homogeneous data, these projects can compare data collected using different protocols and analyzed by different labs. In this way, we can better understand and cope with the variability introduced by collaborative science. Because of the inertia of very large projects, many of the protocols and standards developed for these projects are still in use more broadly. Even though these large-scale efforts can be expensive, they hold great promise for answering big questions in biology. 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 built on to achieve this larger goal. Such efforts should take the lead in the development and dissemination of data and protocol standards across biological and environmental sciences, and on the comparison of results from different protocols and laboratories/groups to develop data transformations needed for integration.

Machine learning and AI for automated knowledge discovery and data management. There is more data integration work to do than what we can possibly afford. Automated solutions will accelerate the work of integrating data across types and scales. Machine learning (ML) and artificial intelligence (AI) can help to determine what is important, fill data gaps, and create metadata. For example, in order to integrate data collected at different spatial scales, downscaling or upscaling is needed, which is the inference of new data using an algorithm. Metadata creation is dominated by semi-automated tools and online templates. We recommend the development and refinement of algorithms for filling gaps in data and metadata. This could include tagging data with ontology terms, spatial and temporal downscaling, and conversion from one format to another. In addition to these algorithms, we must also have proper documentation, provenance, and user interfaces for human-mediated quality assurance. High priority automated tasks include named entity recognition in biological data sets, data typing,

protocol detection, and transformation across formats, methods, and units. Additional investment in community-development work can identify additional high priority tasks.

Math to assess the sufficiency of data. How do you know when you have enough data? Current methods include statistics to assess variation among and between treatments that estimate the replication needed to achieve a significant probability value. Even so, it can be hard to know what is truly important to measure when studying an unknown system. Variables are chosen based on ease of measurement rather than importance. We recommend the development of new data science techniques to characterize signal-to-noise[1] patterns in data and error rates so that we can understand when a data set is complete. These techniques need to be able to operate without simplifying assumptions (e.g., Gaussian statistics, Markovian / memoryless dynamics).

New mathematics for biological data. Algebraic topology, differential geometry, fractal geometry, fractional calculus, high-dimensional (multivariate) statistics, statistical signal processing, and machine learning are successful in analyzing and extracting knowledge from specific homogeneous data streams, yet they fail or face algorithmic and interpretation challenges when dealing with multi-modal, highly-dynamic, unstructured data affected by unknown noise sources or unknown perturbations. From a formal perspective, extracting knowledge from biological data requires rigorous algorithms and conservation principles to identify the unknown unknowns that may govern the dynamics of a biological system (i.e., unknown stimuli that transiently act or perturb parts of a biological system, unmeasured biological variables due to limitations in sensing, or interactions that obscure some variables from monitoring). We recommend the construction of a new mathematical framework capable of mining and analyzing data in all kinds of forms (e.g., combinations of series of events, time series, partially observed graph snapshots, time varying graphs). The new mathematics should not require knowing the entire model *a priori*, but rather be able to discover unknowns from the geometry of the data and structure of the causal predictive model that emerges from time ordered data [7].[2] Developing and inventing new algebraic, differential, and fractal geometry inspired frameworks to be able to analyze multi-modal data streams can not only identify the model structure, but also biological functions over specific scales. In order to deal with unknown degrees of nonlinearity, memory, and heterogeneity specific to biological data, the new mathematical techniques will require new computational techniques for quantifying information transfer, processing, and storage, which allows us not only to describe biological dynamics as new models of computation, but also

[1] Of note, not all noise and not all perturbations can be considered to be of wide sense stationary nature or to follow the norms of Gaussian statistics. Removing such simplifying assumptions calls for the development of new data science techniques in mining data streams and constructing mathematical (causal predictive) models. Often, carefully mining the noise can reveal information about the robustness / fragility or likelihood of a catastrophic event in the evolution of biological systems.

[2] This implies that the new mathematics does not require testing for Markovian or a non-Markovian fit, but rather identifies from the cross-analysis of geometrical dimensions of data which variables require a long-range memory or a short-range memory operator.

enables identification of scale-dependent causal interactions and reconstruction of time-varying, complex networks [17,48] [49]. Within the complex networks/systems framework we need mathematical techniques that go beyond first- and second-order statistics (e.g., degree distribution, assortativity) and be able to estimate the higher-order statistics for defining and quantifying performance envelopes and change points / phase transitions in biological systems [5–7]. Due to stochasticity and the nonequilibrium nature of biological systems evolution, verification and validation of new mathematical models of biological systems should not rely solely only on goodness-of-fit measures, but also on new predictive measures to reproduce specific characteristics, patterns, and trends (e.g., a multi-fractal signature, a specific degree of emergence, robustness, self-organization, and complexity). Lastly, the scientific collective effort should not concentrate on fitting specific mathematical models into biological data, but rather discover new mathematical models, rules, and principles that could help us understand the rules of life and inspire new theoretical foundations for artificial intelligence and computer science.

Professional incentives for sharing and collaboration. An important barrier to large-scale biological data integration is difficulty in getting data from individual researchers. Despite demonstrated benefits [50,51], data sharing is viewed as largely altruistic with little professional reward [52]. Some progress has been made in the form of data journals, data citation guidelines [53–55], and the acceptance of data products as valued research output in some contexts. Even the “twitter famous” Research Parasite Award that honors outstanding secondary analysis is bestowed on the data consumers rather than the providers [56,57]. Unfortunately, much data sharing is still accomplished solely due to the “stick” of funding agency and publisher requirements rather than any “carrot” of professional reward. We recommend the development of a much more robust system for recognizing the work of making data findable and reusable that includes the characteristics below.

- Good data practices and interdisciplinary collaborations need to be recognized for promotion. This is a top-down decision at Universities and Institutions.
- A professional development structure needs to be developed for data curators and information scientists who specialize in data stewardship in order to nurture their careers.
- The funding advantage for researchers who hoard data should be removed. Data that are not shared need to be rendered useless even to the provider.[3]
- Professionalize scientific software development practices and metrics.
- Develop data citation guidelines and metrics that transcend discipline, similar to publication citation and metrics
- Making data FAIR [24] needs to have testable metrics for success [58]

Interdisciplinary collaborations and training. It is nearly impossible to know the specific skills students will need at the end of the 5-10 years it will take to complete their training. Rather than solely imparting skills, students should learn the value of being 1) flexible (adaptable to the

[3] Data that are not properly shared are often useless to the provider, as anyone who has tried to use their own data from several years ago has found out.

unknowns), 2) curious (to discover the unknowns), and 3) brave (to tackle and master the unknowns). This can best be taught by example from faculty by engaging in larger interdisciplinary research projects where everyone interacts, participates, and contributes throughout the entire research process from hypothesis formulation to interpretation of the results within the limitations of each individual study.

Discussion

Addressing the problems of data integration in biology is an ongoing process that is worth the investment. The solutions proposed here will not 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. Not only will we see an increase in the scale and scope of scientific insights, we will also see more reproducible science and improved data quality [59].

Community participation is required for making the decisions needed to implement the solutions to the challenges of biological data integration. Mobilizing the community is hard work and requires sustained effort and funding. Fortunately, we can borrow strategies from disciplines that have successfully developed and maintained community-driven standards [60] and the technology industry has proven methods for the development, management, and maintenance of tools and services that engage users. These successful products and processes took years of sustained effort to develop, many by focused standards development bodies (TDWG [61], GSC [62], ESIP [63], OBOFoundry [64], OGC [65], INCF [66]). Single meetings or publications that involve a limited number of stakeholders will not result in standards adoption. Overly broad standards that apply to too many sub-disciplines or data types will be of limited use. Standards development should be bottom-up with a defined decision-making process and governance plan. Discussion and decisions should be open, transparent, and well documented. It is absolutely essential that practitioners are directly involved in developing their own standards and best practices for data, sharing, and tool development.

In addition to the specific concerns in the above challenges and proposed solutions, there are several considerations that underlie all data integration efforts. Large-scale data integration will inevitably require reuse of data collected outside the laboratory or discipline doing the integration. Thus, the following issues with data reuse are also issues for data integration that transcend the biological sciences. Each of these issues has been discussed in detail elsewhere and have solutions that can be borrowed from other disciplines and adapted for biology. We will briefly review them here.

Data quality and trust

There is no such thing as a perfect data set. Data consumers need to be able to assess the quality of the data they seek to use [22,23,67]. Errors must be reported with feedback and fixed with documentation [68,69]. This is critical for citizen science data, which is invaluable, but can have a real or perceived quality problem [70,71]. In addition to managing human error,

consumers need to trust that data have not been corrupted or altered while being prepared for or stored in a repository [72]. Tools, services, metrics, and best practices have been developed by the library and information science communities for this purpose.

Data ownership and attribution

Intellectual property is very important in science, where ideas are currency. Researchers rightly feel ownership over the data they collect, especially if the data were hard to collect. Nevertheless, data providers are making data available for reuse via a suite of licenses [73] and with a suggested citation [53]. Unfortunately, data citation and licensing are still non-standard, making large-scale reuse and metrics-tracking difficult [74,75]. Integrating data leads to additional challenges with attribution and provenance tracking of the original data sets and the integrated data set. Licensing can add unintended restrictions on reuse; thus, it is always best to use the least restrictive license possible, such as CC-0 [76,77].

Data preservation

Data integration and reuse implies that data are preserved for this purpose; however, there are far more data collected than what is possible to preserve in the long-term. Decisions must be made about what data are worth preserving and for how long. Librarians have been making these decisions for centuries. More recently, the information science community has made recommendations for ways that the biological and environmental science disciplines can prioritize data sets for preservation [78–80].

Data rescue

The vast majority of research output falls within the “long tail of dark data” which is not discoverable or accessible [81]. These data are typically in legacy formats like floppy disks or non-computable formats like paper notes [23]. Transformation of these data at scale is incredibly difficult and resource intensive with few generalizable solutions. Much of these data cannot be re-collected, making their loss particularly concerning. Finding a scalable path for the transformation of legacy data would have a big impact. Efforts within CODATA [82], the Research Data Alliance [83], and the earth sciences community [84] are working to educate researchers about the importance of data rescue and incentivize the task. Unlike the issues discussed above, data rescue has not found a good resolution and still requires significant effort.

The discussion of data integration necessarily includes regret that disciplines, sub-disciplines, and laboratories are not as collaborative or interoperable as they could be. Not just from a data integration perspective, discovery of tools and services is very difficult across disciplinary boundaries. The typical response to seeing value in working across disciplines is to create another discipline that combines elements of the first two (Fig. 3). A large part of the focus on interdisciplinary work has been bringing together practitioners of the disciplines looking to collaborate. This is important, but just as much thought needs to go into the crosswalks between data. Many disciplines and data types have a unifying architecture, such as the periodic table of

the elements or the globe. Rather than solely focusing on how we build relationships between people, let's also invest in building relationships across the data architectures.

An unfortunate consequence of human-centric, cross-disciplinary strategies is that cross disciplinary data and tool discovery is primarily accomplished by word-of-mouth, making it highly reliant on an individual's professional network. If the right people are not in your network you won't find the resources. Discovery cannot solely rely on word-of-mouth and building the data-centric, cross-disciplinary infrastructure is key to having truly interdisciplinary data and tool discovery.

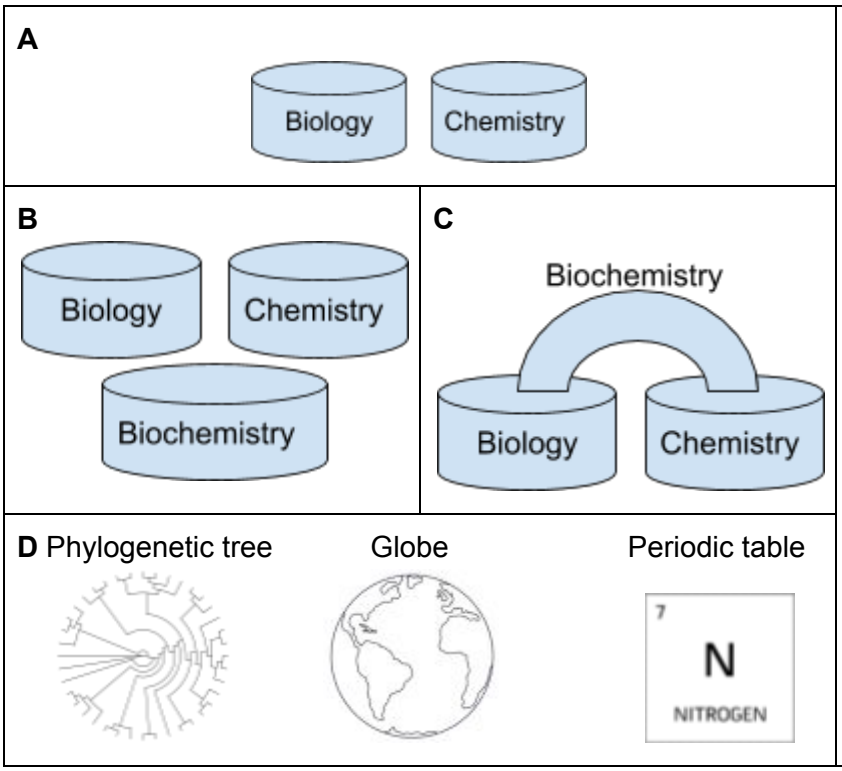


Figure 3. Integrating Disciplines Through Data Infrastructure. Scientific disciplines can be very siloed (A). Much work to promote cross-disciplinary collaboration has focused on expanding the networks of practitioners with less attention paid to integrating data and information architectures. As a result, instead of creating a bridge across disciplinary silos (C), another silo is created (B). The data do not know their discipline. We propose peeling away the disciplinary silos and taking a more data-centric approach, focusing on building bridges across information architectures (D).

Summary

The reintegration of the subdisciplines of biology, and the accompanying insights into the rules of life, require the reintegration of data. Without good data collection and management practices and data science, integration at scale becomes nearly intractable and puts these insights out of reach. Here we outlined key current challenges related to the nature of data, the nature of biological systems and the nature of people, and recommended solutions to these challenges with the aim of reintegrating biology through data. Significant work is required to develop data standards, best practices, and new mathematical approaches, and shift professional incentives that can start to overcome the barriers to data integration. Funding agencies can help by specifically supporting efforts to create community-driven data standards and interdisciplinary data architectures. Universities and Institutions can help by specifically rewarding data sharing and interdisciplinary work. 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.

References

1. Ball A. Review of data management lifecycle models. University of Bath, IDMRC; 2012.
2. Data Life Cycle | DataONE. [cited 18 Dec 2019]. Available: https://www.dataone.org/data-life-cycle
3. Data Lifecycle | NNLM. [cited 18 Dec 2019]. Available: https://nnlm.gov/data/thesaurus/data-lifecycle
4. 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.
5. West BJ. Fractal physiology and the fractional calculus: a perspective. Front Physiol. 2010;1: 12.
6. West BJ. Homeostasis and Gauss statistics: barriers to understanding natural variability. J Eval Clin Pract. 2010;16: 403–408.
7. 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.
8. Maple Ferryman Pty Ltd. In: Maple Ferryman Pty Ltd [Internet]. [cited 18 Dec 2019]. Available: https://www.mapleferryman.com/natures-envelope
9. 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.
10. 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.
11. 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.
12. 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.

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

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

15. 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.

16. 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.

17. Gupta G, Pequito S, Bogdan P. Learning Latent Fractional dynamics with Unknown Unknowns. 2019 American Control Conference (ACC). 2019. doi:10.23919/acc.2019.8815074

18. 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.

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

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

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

22. 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.

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

24. 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.

25. 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.

26. 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.

27. Edwards PN, Mayernik MS, Batcheller AL, Bowker GC, Borgman CL. Science friction: data, metadata, and collaboration. Soc Stud Sci. 2011;41: 667–690.

28. 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.

29. The Carpentries. In: The Carpentries [Internet]. [cited 15 Dec 2019]. Available: https://carpentries.org/

30. Trust and Mistrust in Americans' Views of Scientific Experts. In: Pew Research Center Science & Society [Internet]. 2 Aug 2019 [cited 20 Dec 2019]. Available: https://www.pewresearch.org/science/2019/08/02/trust-and-mistrust-in-americans-views-of-scientific-experts/

31. Updated: What does it take to get your grant targeted by Congress? In: Science | AAAS [Internet]. 12 Nov 2014 [cited 20 Dec 2019]. Available: https://www.sciencemag.org/news/2014/11/updated-what-does-it-take-get-your-grant-targeted-congress

32. Keyes R. The Post-Truth Era: Dishonesty and Deception in Contemporary Life. Macmillan; 2004.

33. 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.

34. 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.

35. ESIP | Connecting Science, Data and Users. [cited 15 Dec 2019]. Available: https://www.esipfed.org/

36. 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.

37. Perez-Harguindeguy N, Diaz S, Garnier E, Lavorel S, Poorter H, Jaureguiberry P, et al. Corrigendum to: new handbook for standardised measurement of plant functional traits worldwide. Aust J Bot. 2016;64: 715–716.

38. Sack L, Cornwell WK, Santiago LS, Barbour MM, Choat B, Evans JR, et al. A unique web resource for physiology, ecology and the environmental sciences: PrometheusWiki. Funct Plant Biol. 2010;37: 687–693.

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

40. BRIDG. In: CDISC [Internet]. [cited 15 Dec 2019]. Available: https://www.cdisc.org/standards/domain-information-module/bridg

41. OMOP Common Data Model – OHDSI. [cited 15 Dec 2019]. Available: https://www.ohdsi.org/data-standardization/the-common-data-model/

42. 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.

43. 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.

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

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

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

47. 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

48. Shalizi CR, Crutchfield JP. Computational Mechanics: Pattern and Prediction, Structure and Simplicity. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.83.8329&rep=rep1&type=pdf

49. Xue Y, Bogdan P. Reconstructing missing complex networks against adversarial interventions. Nat Commun. 2019;10: 1738.

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

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

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

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

54. 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

55. 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

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

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

58. 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

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

60. 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.

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

62. 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.

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

64. Wg OT. The OBO Foundry. [cited 18 Dec 2019]. Available: http://www.obofoundry.org/

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

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

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

68. Morris PJ, Kelly MA, Lowery DB, Macklin JA, Morris RA, Tremonte D, et al. Filtered Push: annotating distributed data for quality control and fitness for use analysis. AGU Fall Meeting Abstracts. 2009. Available: http://adsabs.harvard.edu/abs/2009AGUFMIN34B..08M;

69. Cicero C, Spencer CL, Bloom DA, Guralnick RP, Koo MS, Otegui J, et al. Biodiversity Informatics and Data Quality on a Global Scale1. The Extended Specimen: Emerging Frontiers in Collections-Based Ornithological Research. 2017. Available: https://books.google.com/books?hl=en&lr=&id=cMItDwAAQBAJ&oi=fnd&pg=PT436&dq=data+management+in+github+issues+vertnet&ots=vUYHgiYghG&sig=moNKfFDPfD728YYedmfKUlnlaTU

70. Hunter J, Alabri A, van Ingen C. Assessing the quality and trustworthiness of citizen science data. Concurr Comput. 2013;25: 454–466.

71. Grimm KE. Importance of citizen science for science, individuals, communities, and the planet. Ecology. 2017. pp. 3229–3230. doi:10.1002/ecy.1987

72. Dillo I, Leeuw L de. CoreTrustSeal. Mitteilungen der Vereinigung Österreichischer

Bibliothekarinnen & Bibliothekare. 2018;71: 162–170.

73. Lamy M-H. Creative Commons CC: When we share, everyone wins. 2018. Available: https://creativecommons.org/

74. 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.

75. Robinson-García N, Jiménez-Contreras E, Torres-Salinas D. Analyzing data citation practices using the data citation index. Journal of the Association for Information Science and Technology. 2016;67: 2964–2975.

76. 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.

77. Poisot TE, Mounce R, Gravel D. Moving toward a sustainable ecological science: don't let data go to waste! Ideas Ecol Evol. 2013;6. Available: https://ojs.library.queensu.ca/index.php/IEE/article/view/4632/4992

78. Berman F. Got data?: a guide to data preservation in the information age. Commun ACM. 2008;51: 50–56.

79. Decide what data to preserve | DataONE. [cited 17 Dec 2019]. Available: https://www.dataone.org/best-practices/decide-what-data-preserve

80. Five steps to decide what data to keep | Digital Curation Centre. [cited 17 Dec 2019]. Available: http://www.dcc.ac.uk/resources/how-guides/five-steps-decide-what-data-keep

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

82. Data at Risk - CODATA. [cited 19 Dec 2019]. Available: http://www.codata.org/task-groups/data-at-risk

83. Data Conservation IG. In: RDA [Internet]. 14 Aug 2019 [cited 19 Dec 2019]. Available: https://www.rd-alliance.org/groups/data-conservation-ig

84. Elsevier. International Data Rescue Award in the Geosciences. [cited 15 Dec 2019]. Available: https://www.elsevier.com/awards/international-data-rescue-award-in-the-geosciences