PERSPECTIVE

Challenges and Opportunities of Open Data in Ecology

O. J. Reichman,* Matthew B. Jones, Mark P. Schildhauer

Ecology is a synthetic discipline benefiting from open access to data from the earth, life, and social sciences. Technological challenges exist, however, due to the dispersed and heterogeneous nature of these data. Standardization of methods and development of robust metadata can increase data access but are not sufficient. Reproducibility of analyses is also important, and executable workflows are addressing this issue by capturing data provenance. Sociological challenges, including inadequate rewards for sharing data, must also be resolved. The establishment of well-curated, federated data repositories will provide a means to preserve data while promoting attribution and acknowledgement of its use.

Ecology is an integrative, collaborative discipline (*1*, *2*), amplifying the need for open access to data. The field has rapidly matured over the past century from small-scale, short-term observations and experiments conducted by individuals to include large-scale, long-term, multidisciplinary projects that integrate diverse data sets using sophisticated analytical approaches. Ecological investigations often require interactions with adjacent disciplines (e.g., evolution, genomics, geology, oceanography, and climatology) and disparate fields (e.g., epidemiology and economics). This broad scope generates major challenges for finding effective ways to discover, access, integrate, curate, and analyze the range and volume of relevant information.

The recent *Deepwater Horizon* oil spill in the Gulf of Mexico (*3*) presents a compelling example of the need for far better data access and preservation in ecology and science in general. Understanding spill impacts requires data for benthic, planktonic, and pelagic organisms, chemistry (for oil and dispersants), toxicology, oceanography, and atmospheric science, among others. It also requires data on economic, policy, and legal decisions that affect spill response and cleanup. Despite a few well-organized research groups that can provide relevant data (e.g., the Florida Coastal Ecosystems Long Term Ecological Research site) (*4*), most current and historical data germane to the spill are inaccessible or lost. Furthermore, despite numerous studies associated with past calamities, such as the Ixtoc spill in the Gulf of Mexico (*5*), only a small fraction of the data from these studies is available today. Consequently, our ability to understand both short-term and chronic effects of oil spills is severely limited. As these examples illustrate, access to data is not only important for basic ecological research but also crucial for addressing the profound environmental concerns we face today and, inevitably, in the future.

Unfortunately, only a small fraction of ecological data ever collected is readily discoverable and accessible, much less usable. Based on our own experience building data archives for ecology, we estimate that less than 1% of the ecological data collected is accessible after publication of associated results (*6*, *7*). Rather than providing direct access to data, we share interpretations of distilled data through presentations and publications. To realize advances that are possible through ecological and environmental synthesis, we need to solve the technological and sociological challenges that have limited open access to data. While "open data" will enhance and accelerate scientific advance, there is also a need for "open science"—where not only data but also analyses and methods are preserved, providing better transparency and reproducibility of results.

Solving Technology Challenges

Reviews of ecological informatics have described three major technological challenges: data dispersion, heterogeneity, and provenance (*8*, *9*). Ecosystems and habitats vary across the globe, and data are collected at thousands of locations. Although large quantities of data representing relatively few data sets are typically managed by major research projects, institutes, and agencies, most ecological data are difficult to discover and preserve because they are contained in relatively small data sets dispersed among tens of thousands of independent researchers. Data heterogeneity creates challenges due to the breadth of topics studied by ecologists and the varied experimental

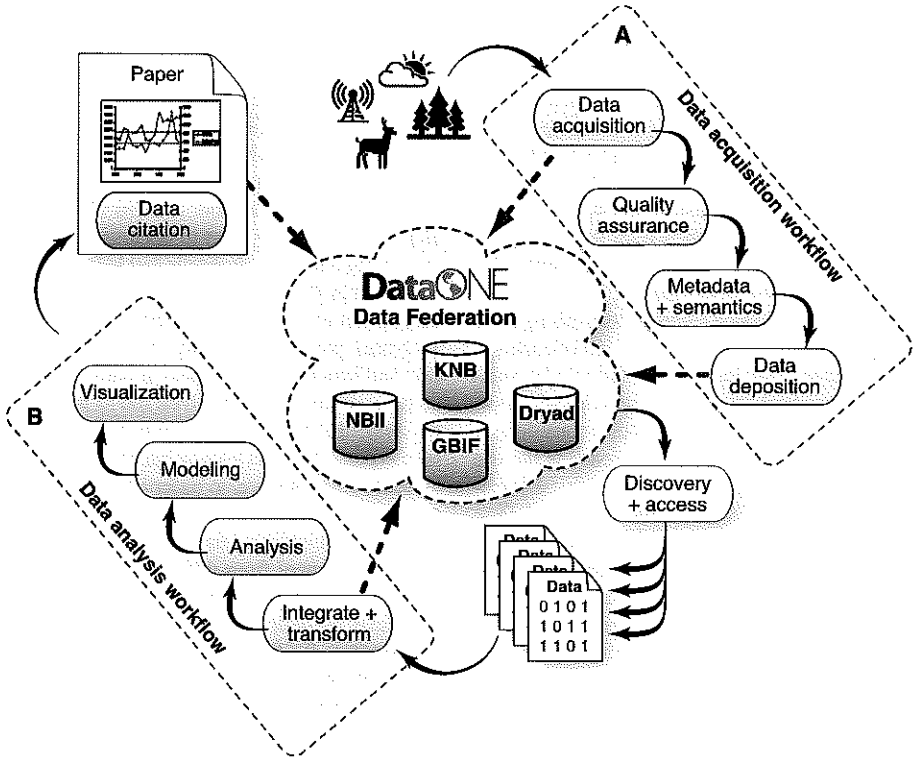

Fig. 1. Data on ecological and environmental systems are **(A)** acquired, checked for quality, documented using an acquisition workflow, and then both the raw and derived data products are versioned and deposited in the DataONE federated data archive (red dashed arrows). Researchers discover and access data from the federation and then **(B)** integrate and process the data in an analysis workflow, resulting in derived data products, visualizations, and scholarly papers that are in turn archived in the data federation (red dashed arrows). Other researchers directly cite any of the versioned data, workflows, and visualizations that are archived in the DataONE federation.

National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara, 735 State Street, Suite 300, Santa Barbara, CA 93101, USA.

*To whom correspondence should be addressed. E-mail: ojreichman@gmail.com

protocols used by independent researchers. Data provenance—origins and history—is necessary when, as is typical in ecological research, interesting results emerge after the data go through complex, multistep processes of aggregation, modeling, and analysis.

The dispersed data issue has been partially addressed by large regional and subject-oriented data collections [e.g., Global Biodiversity Information Facility specimen records (*10*), the Knowledge Network for Biocomplexity (*11*, *12*), the Dryad repository (*13*), and the National Biological Information Infrastructure Metadata Clearinghouse (*14*)]. Unfortunately, these and related efforts are still highly fragmented and collectively have not reached the critical mass of holdings to make them comprehensive. However, several initiatives to federate currently independent data networks are under way. For example, the DataONE (Fig. 1) project is enabling federated access to ecological data from all of the initiatives listed above and creating straightforward mechanisms for new data providers to join the federation. Similar efforts, such as the Data Conservancy project and the international Global Earth Observation System of Systems (GEOSS), will build large federations that eventually will be cross-linked and interoperable with one another and DataONE.

The heterogeneity of ecological data must be addressed when developing technological solutions for managing ecological information. Heterogeneous data in ecology arises from its diversity of subdisciplines (e.g., ecosystems/community ecology, marine/freshwater/terrestrial ecology, and plant/animal/microbial ecology). In addition, adjacent disciplines in earth and life science, as well as relevant disciplines in the social sciences and humanities, have their own terminologies, specialized measurements, and experimental designs that generate heterogeneity.

One way to reduce complications arising from data heterogeneity involves adoption of common experimental practices and measurement standards. Logistical constraints and research priorities often make such an approach impractical. A more generically applicable approach is the use of structured metadata such as Ecological Metadata Language (EML) and the Biological Data Profile that have proven useful for characterizing heterogeneous data. Formal metadata specifications provide guidance for consistently describing data objects and data types (e.g., methods, units of measurement, and details of experimental design). However, these metadata systems provide topically labeled boxes to be filled in using natural language and therefore are not amenable to automated interpretation by computers. Advances in the use of controlled vocabularies (e.g., ontologies) will provide well-defined terms to fill in these boxes and enable computers to more precisely assist a researcher in locating and processing data of interest.

The Semantic Web in particular is beginning to enhance data interoperability. Linked Open Data methods provide ways to connect together data from distributed sites using standard Web technologies, thereby attaching semantic descriptions to data resources (*15*). Unlike most Web sites that only provide human-readable Web pages, linked data allows computers to discover and collate data from the Web without human intervention, enabling new types of synthetic data studies at much larger scales. In addition, unified models for representing the semantics of scientific observations and measurements are emerging within a variety of communities, and these can be used in the linked data cloud. These efforts are useful for representing the semantics of ecological observations and for building tools that directly support synthesis through precise data search and automated data integration (*16*). Although the conceptual basis for these observational data modeling approaches has been demonstrated, substantial implementation must occur before semantic modeling will be available in common ecological data management tools.

Another major challenge is the critical need to track the provenance of derived data objects and scientific results from initial data collection, through quality assurance, analysis, modeling, and ultimately publication (*17*). Provenance is especially important to support scientific results used in policy and management decisions, where field experiments and techniques may not be fully reproducible due to difficulty of replicating environmental conditions.

Computer scientists are making considerable progress in developing ways to capture provenance information. Scripted analysis systems like R, and scientific workflow systems like Kepler and Taverna, can be used to document the data processing and analysis details that led to a given set of results (Fig. 1). Scientific workflow applications can record critical information about the analytical process, including details about the data and how it is transformed, providing a comprehensive record of an analysis and its results. In this way, the data, analytical processes, and results become part of a knowledge base supporting evidence-based science, to better inform decision-making in conservation and resource management (*18*). In addition, new research shows that provenance traces from different studies can be linked when producing synthetic analyses that reuse existing data (*19*). The combination of formal systems for tracking provenance, and federated data repositories like DataONE that provide unique identifiers for every data object, will be instrumental in realizing the goal of fully reproducible science in support of understanding global environmental issues such as climate change, species invasions, and epidemics.

Solving Sociological and Cultural Challenges

Although it is challenging to develop new technological solutions to data sharing in ecology, the social and cultural barriers may be even more onerous. Technical solutions will emerge that considerably enhance access to ecological data, but overcoming the cultural and sociological barriers to increased data access requires changing human behavior.

Some disciplines (e.g., astronomy and oceanography) have a history of sharing data, perhaps because these fields rely on large, shared infrastructure. Other disciplines, such as genomics, also have shared repositories, largely due to the homogeneity of their data. Traditionally, ecologists have had few incentives for sharing information. Research involved gathering and analyzing one's own data and publishing the distilled results in peer-reviewed journals. In addition, sharing data was not viewed as a valuable scholarly endeavor or as an essential part of doing science. Recent advances in ecological synthesis, however, are rapidly changing these attitudes to data sharing.

Researchers might still be disinclined to share their data until they have fully completed analyzing and reporting on their observations and results. The concern is that if data are made openly available in the interim they may be used by other investigators, effectively scooping the data originators. Properly curated data alleviates this concern, as the use of data without permission or attribution would be condemned by colleagues and funding sources. Proper curation requires time and money and is inadequately supported in research funding.

Establishment of a reward system should further motivate investigators to share their data. For example, if data sets are publishable and citable (e.g., *Ecological Archives* and Dryad), they will become more respected and valued as an important part of research and scholarship (*20*). The most effective means to alter the reward system is to make data sharing an expectation of funding and publications and reward those who meet these expectations. The National Science Foundation in the United States now requires an explicit data management plan in all proposals, which is a step in the right direction. Journals and societies that mandate data publication concurrently with research publications also have proven to be effective (e.g., GenBank).

In addition to support for individual researchers to prepare and submit their data to public archives, the community needs to identify sustainable models for federated data archives that persist over decadal time scales. Models such as DataONE involve leveraging institutional contributions in a large federation to protect against uneven funding for individual institutions. Nevertheless, even these initiatives will not work without a sustained commitment from funding agencies that is specifically targeted at institutional data repositories and coordinating organizations.

The evolution of GenBank offers evidence that technological advances and cultural metamorphosis generate paradigm shifts in science.

With a tug from software to manage genomic data online and a push from publishers unwilling to continue editing and printing the growing volume of gene sequences, a robust data repository for gene sequences was born. Today, after almost 30 years, registering gene sequences and sharing them broadly is the norm and is recognized as fostering one of the greatest scientific revolutions in the past century.

Ecology is poised for a similar transformation. The pull comes from a need for data in synthesis and cross-cutting analysis that is facilitated by the emergence of community metadata standards and federated data repositories that span adjacent disciplines. The push is coming from funding entities that are requiring open access to data, with a dose of urgency engendered by the chronic and acute environmental degradation occurring globally. Furthermore, the rewards for sharing data are increasing. As noted, it is possible to publish peer-reviewed, citable data sets in repositories while giving credit to the data contributors, and there is evidence that published papers that do make available their data are cited more frequently than those that do not (*21*).

We have presented some of the major challenges and emerging solutions for dealing with the vast volume and heterogeneity of ecological data. To accelerate the advance of ecological understanding and its application to critical environmental concerns, we must move to the next level of information management by providing revolutionary new data-management applications, promoting their adoption, and hastening the emergence of communities of practice. Concurrently, we must encourage the growing culture of collaboration and synthesis that has emerged in ecology that is fundamentally altering the scientific method to require comprehensive data sharing, as well as greater reproducibility and transparency of the methods and analyses that support scientific insights.

References and Notes

1. S. Carpenter *et al.*, *Bioscience* **59**, 699 (2009).
2. E. Hackett, J. Parker, D. Conz, D. Rhoten, A. Parker, in *Scientific Collaboration on the Internet*, G. M. Olson *et al.*, Eds. (MIT Press, Boston, 2008), pp. 277–296.
3. T. J. Crone, M. Tolstoy, *Science* **330**, 634 (2010).
4. Florida Coastal Everglades Data Resources, http://fce.lternet.edu/data/FCE.
5. J. W. Tunnell, Q. R. Dokken, M. E. Kindinger, L. C. Thebeau, "Effects of the Ixtoc I oil spill on intertidal and subtidal infaunal populations along lower Texas coast barrier island beaches," in *Proceedings of the 1981 Oil Spill Conference* (American Petroleum Institute, Washington, DC, 1981), pp. 467–475.
6. C. J. Savage, A. J. Vickers, C. Mavergames, *PLoS ONE* **4**, e7078 (2009).
7. P. B. Heidorn, *Libr. Trends* **57**, 280 (2008).
8. W. K. Michener, *Ecol. Inform.* **1**, 3 (2006).
9. M. B. Jones, M. Schildhauer, O. J. Reichman, S. Bowers, *Annu. Rev. Ecol. Evol. Syst.* **37**, 519 (2006).
10. V. S. Chavan *et al.*, *State-of-the-Network 2010: Discovery and Publishing of the Primary Biodiversity Data Through the GBIF Network* (Global Biodiversity Information Facility, Copenhagen, 2010).
11. S. J. Andelman, C. Bowles, M. R. Willig, R. Waide, *Bioscience* **54**, 240 (2004).
12. The Knowledge Network for Biocomplexity, http://knb.ecoinformatics.org.
13. H. White, S. Carrier, A. Thompson, J. Greenberg, R. Scherle, "The Dryad data repository: A Singapore framework metadata architecture in a DSpace environment," in *Proceedings of the International Conference on Dublin Core and Metadata Applications*, J. Greenberg, W. Klas, Eds. (Dublin Core Metadata Initiative and Universitätsverlag Göttingen, Berlin, 2008), pp. 157–162.
14. I. San Gil, V. Hutchison, G. Palanisamy, M. Frame, *J. Libr. Metadata* **10**, 99 (2010).
15. C. Bizer, T. Heath, K. Idehen, T. Berners-Lee, "Linked data on the Web (LDOW2008)," in *Proceedings of the 17th International Conference on World Wide Web* (Association for Computing Machinery, New York, 2008), pp. 1265–1266.
16. J. Madin, J. S. Bowers, M. Schildhauer, M. B. Jones, *Trends Ecol. Evol.* **23**, 159 (2008).
17. P. Buneman, S. Khanna, W.-C. Tan, *Lect. Notes Comput. Sci.* **1974**, 87 (2000).
18. W. Sutherland, A. Pullin, P. Dolman, T. Knight, *Trends Ecol. Evol.* **19**, 305 (2004).
19. P. Missier *et al.*, Linking Multiple Workflow Provenance Traces for Interoperable Collaborative Science., Presentation at WORKS 2010: 5th Workshop on Workflows in Support of Large-Scale Science, IEEE Computer Society, New Orleans, 14 November 2010.
20. T. Vision, *Bioscience* **60**, 330 (2010).
21. H. A. Piwowar, R. S. Day, D. B. Fridsma, J. Ioannidis, *PLoS ONE* **2**, e308 (2007).

22. Supported by the National Center for Ecological Analysis and Synthesis, a Center funded by NSF (grant EF-0553768), the University of California, Santa Barbara, and the State of California. Additional support for M.B.J. was provided by NSF grant OCI-0830944 and for O.J.R. by NSF grant DEB-0444217.

10.1126/science.1197962

PERSPECTIVE

Changing the Equation on Scientific Data Visualization

Peter Fox and James Hendler*

An essential facet of the data deluge is the need for different types of users to apply visualizations to understand how data analyses and queries relate to each other. Unfortunately, visualization too often becomes an end product of scientific analysis, rather than an exploration tool that scientists can use throughout the research life cycle. However, new database technologies, coupled with emerging Web-based technologies, may hold the key to lowering the cost of visualization generation and allow it to become a more integral part of the scientific process.

A critical aspect of the data deluge is the need for users, whether they are scientists themselves, funders of science, or the concerned public, to be able to discover the relations among and between the results of data analyses and queries. Unfortunately, the creation of visualizations for complex data remains more of an art form than an easily conducted practice. What's more, especially for big science, the resource cost of creating useful visualizations is increasing: Although it was recently assumed that data-centric science required a rough split between the time to generate, analyze, and publish data (*1*), today the visualization and analysis component has become a bottleneck, requiring considerably more of the overall effort. This trend will continue to get worse as new technologies for data generation are decreasing in price at an incredible rate (in terms of cost per data generated), whereas visualization costs are falling much more slowly. As a result of these trends, the extra effort of making our data understandable, something that should be routine, is consuming considerable resources that could be used for many other purposes.

A consequence of the major effort for visualization is that it becomes an end product of scientific analysis, rather than an exploration tool allowing scientists to form better hypotheses in the continually more data-intensive scientific process. However, new database technologies and promising Web-based visualization approaches may be vital for reducing the cost of visualization generation and allowing it to become a central piece of the scientific process. As an anecdotal example, consider the papers in the recently published *The Fourth Paradigm*, a collection of invited essays about the emerging area of data-intensive science (*2*). Only one of the more than 30 papers is primarily about visualization needs, but virtually all of the essays include visualizations that show off particular scientific results.

From Presentation

In the computing sciences, visualization has been in the hands of two communities. The first is the

Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY 12180, USA.

*To whom correspondence should be addressed. E-mail: hendler@cs.rpi.edu