Big data and the future of ecology

Stephanie E Hampton[1*], Carly A Strasser[2], Joshua J Tewksbury[3], Wendy K Gram[4], Amber E Budden[5], Archer L Batcheller[6], Clifford S Duke[7], and John H Porter[8]

The need for sound ecological science has escalated alongside the rise of the information age and "big data" across all sectors of society. Big data generally refer to massive volumes of data not readily handled by the usual data tools and practices and present unprecedented opportunities for advancing science and informing resource management through data-intensive approaches. The era of big data need not be propelled only by "big science" – the term used to describe large-scale efforts that have had mixed success in the individual-driven culture of ecology. Collectively, ecologists already have big data to bolster the scientific effort – a large volume of distributed, high-value information – but many simply fail to contribute. We encourage ecologists to join the larger scientific community in global initiatives to address major scientific and societal problems by bringing their distributed data to the table and harnessing its collective power. The scientists who contribute such information will be at the forefront of socially relevant science – but will they be ecologists?

Front Ecol Environ 2013; 11(3): 156–162, doi:10.1890/120103 (published online 12 Mar 2013)

In the 21st century, biology is running full tilt into the information age (Spengler 2000); leaders in many fields of the life sciences, including genomics, nanobiology, and medicine, have embraced the new opportunities presented by unprecedented access to digital information. Global-scale environmental issues, from climate change and food security to the spread of disease and the availability of clean water, are creating pressure for ecologists to collectively step forward into this new age. Society is asking ecologists for information that is both specific to particular problems, places, and times, and also predictive, prescriptive, and scalable.

This is a challenge ecologists cannot meet individually.

In a nutshell:

- Ecologists collectively produce large volumes of data through diverse individual projects but lack a culture of data curation and sharing, so that ecological data are missing from the landscape of data-intensive science
- To fully take advantage of scientific opportunities available in the information age, ecologists must treat data as an enduring product of research and not just as a precursor to publications
- Forward-thinking ecologists will organize and archive data for posterity, publicly share their data, and participate in collaborations that address large-scale questions

[1]*National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara, Santa Barbara, CA* [*]*(hampton@nceas.ucsb.edu);* [2]*California Digital Library, University of California Office of the President, Oakland, CA;* [3]*Department of Biological Sciences, University of Washington, Seattle, WA;* [4]*National Ecological Observatory Network, Boulder, CO;* [5]*DataONE, University of New Mexico, Albuquerque, NM;* [6]*School of Information, University of Michigan, Ann Arbor, MI;* [7]*Ecological Society of America, Washington, DC;* [8]*Department of Environmental Sciences, University of Virginia, Charlottesville, VA*

Our ability to produce specific analytical information for local problems that can also address questions at larger spatial scales and over longer time frames depends on our willingness to work collaboratively to collect, preserve, and share our data across projects, locations, and research groups (Palmer *et al.* 2005). Major public and private investments in data-intensive science have proliferated in recognition of the expanding possibilities for scientific discovery inherent in the rise of "big data" – the increasing volume, variety, and velocity of data streams across sectors. In the US, such investments include the recent White House Office of Science and Technology Policy's funding of big-data initiatives offering scientists many new opportunities to address problems across scales that would otherwise be impossible for individual scientists. The investigators and disciplines that can quickly bring relevant data to bear on complex environmental problems will set the agenda and drive progress.

This increasing societal emphasis on big data presents a problem for our field: collectively, ecologists produce a tremendous amount of data, but ecology has yet to develop a culture of transparent data exchange and aggregation (Jones *et al.* 2006; Ellison 2010; Reichman *et al.* 2011). Ecology is dominated by what Heidorn (2008) has dubbed "long tail" science – science conducted by individual investigators, often over limited spatial and temporal scales, under funding models that provide limited capacity for data curation or sharing. These issues are compounded by a lack of incentives for collaborative data sharing, the sense that ecological data are difficult to understand out of their original context, and a tremendous heterogeneity of ecological data types – from arduous behavioral observations carried out at remote field sites to continuous data streams pouring in from sensor networks (Jones *et al.* 2006). But these challenges are not unique to ecology; researchers in an assortment of scien-

tific disciplines are grappling with similar issues regarding data preservation and sharing (Borgman 2009; Costello 2009) and the complexity of the endeavor has been exacerbated by a proliferation of technology for handling heterogeneous data. We are in an age of data-intensive science and big data, and ecologists must develop the capabilities to deal with their data. Our capacity to efficiently provide timely information to meet modern societal challenges will depend on a global "greening" of ecology – that is, data should not only be generated and analyzed, but must also be available to be re-used and recycled.

Is "big science" the answer? Not entirely

The increased need for ecological information to address major environmental problems has led to multiple calls for a rapid shift to "big science" ecology (Costello 2009; Kelling *et al.* 2009; Aronova *et al.* 2010; Ellison 2010). The term "big science" was coined in 1961 by Alvin Weinberg, then director of the Oak Ridge National Laboratory (ORNL), to describe the large, complex scientific endeavors in which society makes sizeable investments, often in the form of government funding (Weinberg 1961). These projects tend to involve international, collaborative efforts among many scientists and institutions (Price 1963; Borgman *et al.* 2007; Aronova *et al.* 2010) and are often characterized by expensive shared equipment (Borgman *et al.* 2007).

Toward the middle of the past century, the International Geophysical Year (IGY; 1957–1958) marked the entry of big science in academic, data-driven research – in this context, big science is defined as broadly inclusive scientific collaboration, with organizational infrastructure for large-scale research (Aronova *et al.* 2010). The IGY was considered a great success, with important outcomes that shaped the field of geophysics and altered the course of science (Kwa 1987). In an effort to extend the idea of big science to ecology, the International Biological Program (IBP) was launched in 1964, shortly after the IGY ended, and ran until 1974.

The story of the IBP is a case study of ecology's uneasy relationship with big science. From the outset, the organizing group had trouble agreeing on a unifying conceptual framework (Aronova *et al.* 2010). They resisted the idea of worldwide cooperative research in ecology because this implied the centralization and homogeneity of research methods and approaches, concepts not inherent in traditional ecology (Michener *et al.* 2007). Rather than embracing the idea of big science and accepting its challenges, many ecologists instead considered it to be a contagion that should be prevented from spreading to ecology. As a result, the IBP accommodated many small-scale investigations and local initiatives, rather than the large collaborative projects that were initially envisioned, and received less centralized funding than the IGY. The major legacy of the IBP was

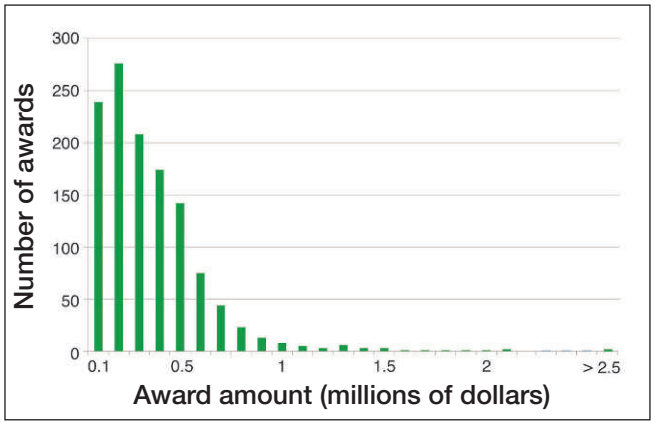

Figure 1. *Number of NSF Division of Environmental Biology awards between 2005 and 2010 (*n *= 1234) binned by award size (mean = $311 800, ± $8421 standard error). Awards for dissertation improvement grants, workshops, and symposia were excluded.*

the Biome Analysis-of-Ecosystems program, which used a systems ecology approach to understand particular biomes in the US, and emphasized sharing data and resources to facilitate rapid progress on research questions (Kwa 1987; Aronova *et al.* 2010). At the end of the IBP, the National Science Foundation (NSF) approved a continuation of this approach in the form of the Long Term Ecological Research (LTER) Program, which officially started in 1980. The LTER Program, which continues today, represents an interesting hybrid between "small" science and big science – individual scientists and small teams of researchers work on a series of problems in targeted areas (eg understanding processes and patterns associated with primary productivity or disturbance) but within the framework of a long-term project that has the potential to document patterns over much larger spatial and temporal scales (Aronova *et al.* 2010).

After the initial struggles of the IBP, the ecological community scaled back attempts to carry out big science, instead continuing to primarily pursue investigator-driven research, which remains the predominant model for ecological research. This does not, however, prevent us from generating big data en masse. In fact, it could be argued that ecologists are already collectively producing big data – a high volume of high-value data – but we are not harnessing its power.

Traditional ecology produces "dark data"

Across the sciences that are supported by funding from the NSF, 2% of the largest awards in 2007 accounted for 20% of the total budget for research (Heidorn 2008). The distribution of funding for environmental science is similar (Figure 1). This distribution means that the vast majority of scientific projects are relatively small; they are not producing big data from supercolliders or satellites, but, when taken together, the amount of data produced

represents a substantial portion of US scientific output.

There are good reasons why the funding is structured in this way. A diverse portfolio of relatively small investments creates a breeding ground for new ideas (Heidorn 2008). Furthermore, smaller projects tend to have higher levels of direct investigator involvement in the data collection, as compared with the automation required for big-science projects to operate effectively. Investigators on smaller projects develop insights through hands-on experiences, resulting in extensive knowledge of study systems and processes that drive observed patterns, while also allowing for the serendipitous discoveries that push forward the frontiers of science (Dunbar 1995). This direct involvement of highly trained researchers in data collection contributes not only to the comparatively high value of each data point but also to a strong sense of ownership (Zimmerman 2003).

This individualistic research approach imposes challenges on the scientific endeavor at large. Heidorn (2008) made the important point that the data produced by a multitude of smaller projects are frequently less available and less well curated than those produced by major initiatives, with less funding and personnel time dedicated to information management. The fate of these smaller datasets is mostly unknown. An examination of data availability from ecological projects (Panel 1) demonstrates what ecologists already intuitively know – they are not making their data available. A vast pool of "dark data" thus constitutes one of the major outputs of US science: potentially invaluable information is largely inaccessible, with only a portion of the data visible through the investigator's publications.

The problem of dark data is one of lost opportunities – what could have been produced had the data been available to others – as well as the additional costs of unnecessary data replication. Diverse, repurposed data can support synthetic approaches at larger scales than was originally intended (Palmer *et al.* 2005; Jones *et al.* 2006), lending new perspectives to old questions and inspiring new lines of inquiry. Funders are becoming increasingly aware of this wastefulness in the use of their research dollars and are encouraging reform (eg www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp).

Much time, energy, and journal space have been dedicated to examining the obstacles and attitudes that have prevented ecologists from making their data publicly accessible (WebPanel 1). These obstacles are real. So why should individual ecologists be motivated to make their everyday data a functional part of ecology's big data in spite of these impediments?

Simply put, the era of data-intensive science is here. Those who step up to address major environmental challenges will leverage their expertise by leveraging their data. Those who do not run the risk of becoming scientifically irrelevant.

Panel 1. Searching for available data from federally funded ecology projects

To ascertain the state of data sharing in ecology, we conducted a survey of ecological papers to determine how much data were publicly available. When designing the survey, our goal was to take an approach that would constitute a "reasonable effort" by an individual ecologist to locate data held by an author with whom the seeker is not acquainted. More details are provided in WebPanel 2.

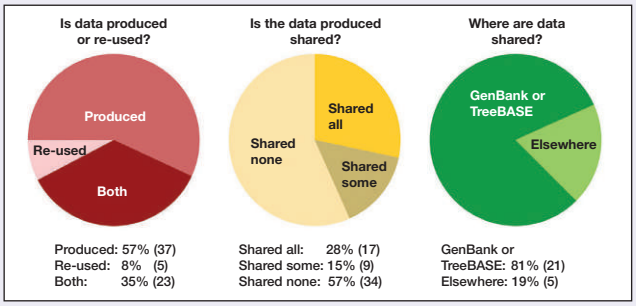

Figure 2. *Results of searching for public data associated with 100 randomly selected papers produced by projects funded through NSF's Division of Environmental Biology (2005–2009).*

We surveyed 100 NSF awards within the NSF's Division of Environmental Biology (DEB), by randomly choosing 20 awards each from years 2005–2009. We randomly selected one paper from each of the awards and assessed (1) what type of paper it was, (2) the data used in the paper, and (3) whether any data from the paper were shared. Papers were categorized as either "using" data or "not using" data. Examples of papers that did not use data were review papers, some model papers, and taxonomy papers. Model papers sometimes re-used data for parameterization, and taxonomy papers sometimes re-used data for creating phylogenetic trees but, for the purposes of our study, effectively no data were produced in the papers put into these categories. Papers that did use data were categorized as producing, re-using, or both. Of those that produced data, we documented whether they shared all, some, or none of the data produced. Sharing *all data* was defined as sharing data necessary for creating all of the primary results of the paper; not all data produced by the study were likely shared in every instance. Sharing some data indicates that there were obvious missing datasets that were used in the study to produce the results reported.

Ecological data are not typically made publically available (Figure 2). We found that of the papers that produced data, only 43% shared some or all of those data. Of those papers that did share data, 81% of that distribution took place through either GenBank or TreeBASE, both of which are associated with genetic data and analyses. Thus, only 8% of papers funded by the DEB made public any of their non-genetic data.

The argument is often made that genetic data are shared because journals *require* that authors provide accession numbers for GenBank or TreeBASE at the time of article submission. To examine this possibility, we assessed journal data-sharing requirements for all the articles in our study that shared data. Of the 48 journals that published papers in which the associated data were shared, 17 contained language stating that authors *must* share their data, with nine specifically mentioning genetic data. What we can conclude from this result is that a cultural shift has occurred in disciplines that produce genetic data; regardless of journal requirements, these data are shared, while other types of data are not.

Panel 2. Citizen science – crowd-sourcing big data for ecology

Every day, people all over the world enter information about birds that they have seen into eBird, a real-time, online checklist program. These "citizen scientists" have enthusiastically shared and archived their data on bird species occurrence and abundance. Using eBird data in association with remote-sensing information on habitat, climate, human population, and demographics, researchers have developed animated maps that show predicted presence or absence of species through time, at finer scales than previously possible. Eventually, these models of bird migratory activity will combine with climate-change scenarios to potentially predict migratory changes for different species (NABCI 2011).

Citizen science is a form of data sharing, and so the challenges of using citizen-science data reflect classic data-sharing challenges more generally. How can one know whether to trust the data? How does one understand the context in which the data were produced? Ecologists have found that dealing with the scale and types of biases present in citizen-science data "has necessitated development of new, more sophisticated approaches to the analysis of large datasets" (Dickinson *et al.* 2010). In many ways, the tools that are helpful for analyzing, visualizing, and sharing citizen-science data complement tools that scientists are developing and using to work with other heterogeneous ecological data (eg Kelling *et al.* 2009).

Citizen-science projects like eBird demonstrate the value of sharing small, localized observations that, when aggregated, build a deeper and broader understanding of ecological phenomena. These data are often publicly available (27 of 53 citizen-science projects that self-categorized themselves as "research based" on the Cornell Lab of Ornithology Citizen Science Toolkit website made their data publicly available at the time we searched the website on April 7–10, 2011) and increasingly used to answer important ecological questions (eg Kelling *et al.* 2009). If volunteers' data have proved so valuable, surely equivalent or greater benefits could be gained by sharing and integrating the data generated by professional scientists. When ecologists choose not to share their data, then researchers, policy makers, and scientists must find other information to address environmental questions at hand, whether or not the data are detailed enough or even appropriate. Time-sensitive natural resource management decisions frequently cannot await a lengthy exchange with researchers to unearth the necessary data, and managers are likely to look toward more readily available public resources, such as citizen-science data; note the utility of eBird data in estimating potential dangers to shore breeding birds in the days immediately following the *Deepwater Horizon* oil well blowout in the Gulf of Mexico (http://ebird.org/content/ebird/news/ebird-gulf-coast-oil-spill-bird-tracker).

Cultural and technological changes in ecology

Ecology is becoming broader, more integrative, and more reliant on large data repositories and automated data collection, although not all of this progress is coming from the "traditional" ecology community. For example, researchers working on many new initiatives are taking advantage of the opportunity to gather environmental data via "citizen scientists" at scales that professional scientists would have difficulty attaining (Panel 2). Continental and regional scales of data collection and public release are planned (Panel 3) through the National Ecological Observatory Network (NEON), and publicly funded sensor networks like the US Integrated Ocean Observing System (IOOS; Baptista *et al.* 2008). These initiatives and observatories provide big-science ecological and environmental data that can be used by anyone, including individuals who are not formally trained as scientists.

Other disciplines, more familiar with using publicly available data resources, are also increasingly addressing important ecological questions. Consider the growing involvement of geophysical sciences in environmental issues, using remote sensing to make inferences about biological processes and patterns that were once solidly within ecology's scope. For instance, the ecological concept of "phenology" is rapidly becoming a common topic at the American Geophysical Union's (AGU's) annual meeting. Comparing prevalence of the word "phenology" between AGU and the Ecological Society

Panel 3. The rise of environmental observing systems

Ecological observatories have arrived (Hamilton *et al.* 2007; Keller *et al.* 2008). In the US, the NSF has committed $434 million to build NEON, with the explicit expectation that data will be made public. These efforts stretch beyond the US; for example, the Terrestrial Ecosystem Research Network is an Australian program designed to capture coordinated monitoring data (www.tern.org.au). The US IOOS is being designed as a marine counterpart (Baptista *et al.* 2008). Naturally, these observatories have had to choose locations and focal measurements very carefully; they cannot be everywhere, monitoring everything at once. Individual investigators will continue to play a critical role in working with organisms, ecosystems, processes, and methods that the observatories do not explicitly address, and in bringing diverse perspectives, analytical approaches, and complementary data to bear on understanding patterns disclosed by observatory data. Ecologists soon will be challenged to modify the individualist approach so that it exists to complement, and preferably to be in synergy with, observatory research.

An undesirable alternative would be that observatory data might stand alone as the only readily discoverable data when pressing societal needs arise, despite the broad acknowledgement that observatories cannot tell the whole story alone. We think this scenario is unlikely, given Darwinian principles of the scientific enterprise; we believe that there will be some forward-thinking ecologists whose datasets are well documented and discoverable online, complete with machine-readable metadata that make data more readily integrated with the massive observatory data streams. Scientists who stand ready with their data to "plug and play" with large datasets are at an advantage relative to those ecologists adhering to data practices that allow their hard-won data to diminish in obscurity.

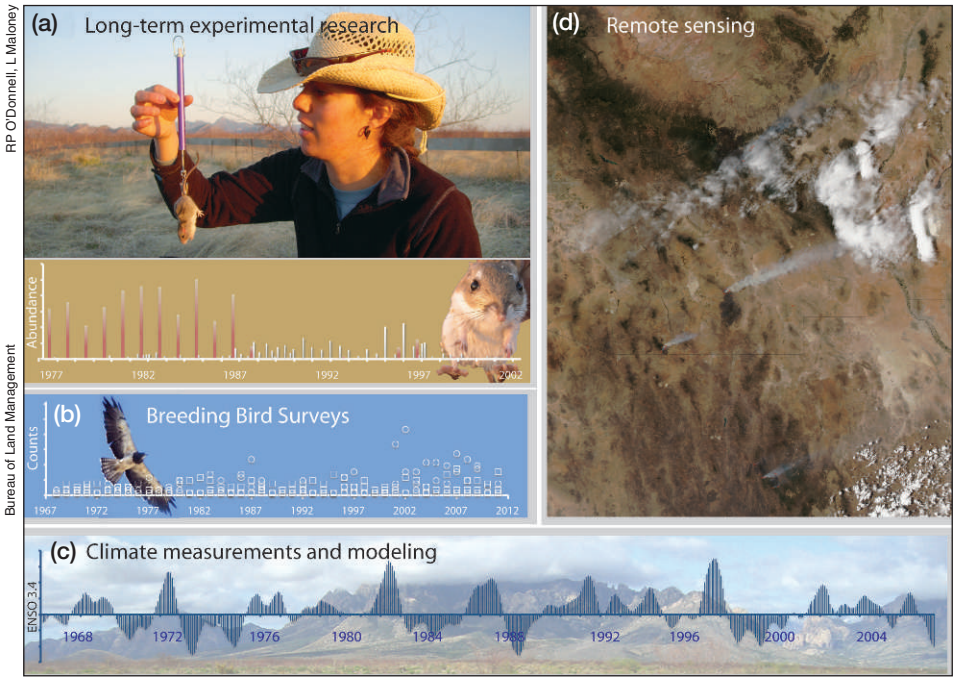

Figure 3. *Publication and robust archive of a 25-year ecological dataset collected near Portal, Arizona (Ernest* et al. *2009) have protected individual and societal investments in long-term ecological research and created possibilities for larger scale data integrations. These public data, such as (a) rodent abundances in experimental kangaroo rat removal plots, have been cited in various studies that originated both inside and outside the original research group. Concurrently, long-term citizen-science efforts in the region, such as (b) the nearby annual Breeding Bird Surveys documenting hawk dynamics, may provide useful data for integration (data from USGS Breeding Bird Survey, www.pwrc.usgs.gov/bbs/RawData). (c) Large-scale climate measurements and modeling provide public data that describe and predict local weather patterns affecting ecosystem dynamics (ENSO 3.4 data from www.cgd.ucar.edu/cas/catalog/climind/TNI_N34). (d) Remote-sensing products capture large-scale landscape patterns, placing individual researchers' work into larger spatial and temporal contexts (Arizona's 2011 Wallow, Horseshoe2, and Monument fires from National Aeronautics and Space Administration, http://rapidfire.sci.gsfc.nasa.gov). As environmental observatories go online (Panel 3), the Portal research site is in a particularly good position to be integrated with NEON data from a new core site less than 300 km away.*

of America's (ESA's) abstracts from 2009 and 2010, 40% of the abstracts related to phenology were presented at AGU meetings.

Many ecologists already work across multiple scales. Those involved in collaborations that share and integrate individually collected, diverse, large-scale data (eg remote-sensing or citizen-science data) are poised not only to gain new scientific perspectives but also to demonstrate that ecological data and expertise are important in addressing major environmental questions. Ecologists who participate in such collaborations can communicate the ecological processes that generate patterns in larger scale data, and the "details" of natural history that are not otherwise evident in imagery or other coarse-scale data. The fine-scale data that ecologists are uniquely positioned to provide and integrate with remote-sensing and other large-scale datasets, including but not limited to the activity of "groundtruthing", will be a critical component of employing these technologies to make new, transformative discoveries at larger scales.

To participate in the information age, ecologists must have well-documented data in standardized formats with standardized, machine-readable metadata, and they must make their data (or at least the metadata) publicly accessible (Figure 3). Ecologists should be prepared to integrate their data with big data streams as well as with more traditional ecological datasets. They need to build the bridge between individual ecology projects and big data and should be the first to respond when funders and journals require data sharing.

Moving into the information age

Examples of successes in sharing and collaboratively integrating ecological data are abundant among the scientific contributions that have emerged from networks of individual ecologists who have teamed up to analyze and synthesize disparate, highly heterogeneous datasets. The integration of scientific results from multiple research projects through meta-analysis is becoming more common in ecology (Chaudhary *et al.* 2010) and has had an increasing impact on the ecological literature over the past 10 years (Cadotte *et al.* 2012). Rigorous quantitative syntheses of existing data have been the sole focus of several research institutes, such as the National Center for Ecological Analysis and Synthesis (NCEAS) since 1995 and the National Evolutionary Synthesis Center (NESCent) since 2004. NCEAS products provide evidence to the individual researcher that thinking big with small datasets is worth the trouble. For example, within 10 years of its establishment, NCEAS had risen to the top 1% of 38 000 institutions worldwide that publish in ecology and environmental science in terms of scientific impact; citation rates for its publications are substantially higher than those of the top ecological journals (Hampton and Parker 2011). Grassroots networks have also been developed outside of formal structures; for instance, the Nutrient Network shares and integrates not only existing data but also new data as it continues to grow (Stokstad 2011).

As networks of scientists develop, advances have also

been made in constructing infrastructure for data archiving and management. Examples include the Global Biodiversity Information Facility, which focuses on making biodiversity data available to the broader community, including the data and synergistic work of many of its partners (eg VertNet, FishBase); the US Geological Survey's (USGS's) National Geological and Geophysical Data Preservation Program, which is tasked with creating a network of geoscience information and data repositories; and the USGS Biodiversity Information Serving Our Nation program, a new integrated resource for biological occurrence data. In addition to programs aimed at facilitating the discovery and preservation of data, there are numerous repositories already in place that house ecological and environmental data (eg the Knowledge Network for Biocomplexity, ORNL's Distributed Active Archive Center, ESA's *Ecological Archives*, iPlant, NatureServe, Dryad, the National Oceanographic Data Center). Some of these repositories house data produced only by affiliated researchers, while others are open to the research community at large. Access to such repositories is critical to researchers in the information age; these facilities and their staff provide stable archiving capabilities for data, enhance data discovery, and increase the visibility of researchers' work. Citation rates of biomedical research articles are boosted by 69% when detailed data underlying the article are made public via a trusted repository (Piwowar *et al.* 2007).

The NSF has recently demonstrated its interest in the preservation and management of data by instituting new regulations that require all submitted research proposals to include data management plans and by establishing the DataNet program, which calls for proposals for "sustainable digital data preservation and access network partners". Five DataNet projects have been funded so far, with DataONE focused chiefly on archiving environmental data and improving interoperability among the existing environmental data repositories (Michener *et al.* 2011). Ecologists would be wise to take note of this shift in funder focus and be ready to address issues related to data sharing, re-use, and archiving.

Other scientific fields are several steps ahead of ecology in embracing the era of big data, and their successes clearly show that individual ecologists could achieve more if they changed their practices toward more open models of research and began treating data as a scientific product of enduring value. For example, in areas of biology that use genetic data, publishing data is now the norm. In the late 1970s, researchers working with nucleic acid sequences recognized a need for safe communal archiving – GenBank and its international partner repositories grew out of these efforts (Strasser 2008). The leading journals in the field also required data deposition to accompany publications based on sequence data, and compliance has been high (85–97%; Noor *et al.* 2006). There is some suggestion that this data-sharing requirement precipitated not only a change in sharing behavior among authors who publish in those journals, but that the culture has shifted to the point where authors now publish their data in appropriate public repositories even when not required to do so (Panel 1). The use and re-use of existing data in these fields is so systemic (Strasser 2008) that it is not possible to estimate how radically data publication and integration have advanced scientific discovery.

Action items for individual ecologists

Ecologists need to treat data as an enduring product of research, not just a precursor to publication. Individual ecologists therefore must:

(1) Organize, document, and preserve data for posterity. Taking data management seriously now will prepare the individual researcher for the time when the incentives are there to integrate data with larger efforts or simply to share data with colleagues and the public. Free software tools are available to produce standardized, machine-readable metadata (eg Morpho, an open-source, spreadsheet-style desktop application that writes Ecological Metadata Language and helps to enforce best practices with data).
(2) Share data. Data federations, such as DataONE, provide linkages among specialized environmental data holdings; in addition, many ecologists have mechanisms for publishing their data through their university libraries, professional journals (eg ESA's *Ecological Archives*, Dryad-associated journals), or other institutions.
(3) Collaborate with networks of colleagues to bring together heterogeneous datasets to address larger scale questions. Ecologists work at a variety of scales that, when integrated, can help to link process and pattern at broad temporal and spatial scales.
(4) Address data management issues with students and peers. Encourage participation in professional workshops, develop data protocols for laboratories and projects, and feature data management in courses through hands-on activities and group discussions – Borer *et al.* (2009) provided a simple introduction to best practices in data management.

Conclusions

Ecology can make critical contributions to large-scale environmental questions and close many knowledge gaps that are likely to persist in big-science initiatives, but only if ecologists are willing to participate in the big-data landscape. For example, ecologists work with the multi-scale data that are needed to supplement the relatively coarse-scale patterns seen in satellite data; ecologists study a wide range of organisms, locations, processes, and methods, covering broader topics relative to observatories; ecologists also have expertise that amateur naturalists seek when they participate in citizen-science initiatives.

Even the smallest datasets can contribute key knowledge for large-scale problem solving, as these data are frequently produced by hands-on work at scales not undertaken by others. A dataset documenting population changes for an endangered species is invaluable to a natural resource manager, but that small dataset is of far less value if its metadata are incomplete, and it is of no value at all if the dataset is never discovered.

Ecologists who thrive in the shifting landscape of the information age will be those who recognize that leveraging our expertise requires us to share our data. These ecologists will treat data as important products of research, bringing ecology into an era of data-intensive research.

Acknowledgements

The DataONE Community Engagement and Education Working Group was supported by the NSF's DataONE award (OCI 0830944). The manuscript was improved by comments from M Jones, S Katz, B Michener, M Schildhauer, and E White. M Ernest generously provided photos and valuable information that contributed to Figure 3.

References

Aronova E, Baker KS, and Oreskes N. 2010. Big science and big data in biology: from the International Geophysical Year through the International Biological Program to the Long Term Ecological Research (LTER) network, 1957–present. *Hist Stud Nat Sci* **40**: 183–224.

Baptista A, Howe B, Freire J, *et al.* 2008. Scientific exploration in the era of ocean observatories. *Comp Sci Eng* **10**: 53–58.

Borer ET, Seabloom EW, Jones MB, and Schildhauer M. 2009. Some simple guidelines for effective data management. *Bull Ecol Soc Amer* **90**: 205–14.

Borgman CL. 2009. The digital future is now: a call to action for the humanities. *Digital Humanities Quarterly* **3**: n4.

Borgman CL, Wallis J, and Enyedy N. 2007. Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *Int J Dig Lib* **7**: 17–30.

Cadotte MW, Mehrkens LR, and Menge DNL. 2012. Gauging the impact of meta-analysis on ecology. *Evol Ecol* **26**: 1153–67.

Chaudhary VB, Walters LL, Bever JD, *et al.* 2010. Advancing synthetic ecology: a database system to facilitate complex ecological meta-analyses. *Bull Ecol Soc Amer* **91**: 235–43.

Costello MJ. 2009. Motivating online publication of data. *BioScience* **59**: 418–27.

Dickinson JL, Zuckerberg B, and Bonter DN. 2010. Citizen science as an ecological research tool: challenges and benefits. *Annu Rev Ecol Evol S* **41**: 149–72.

Dunbar K. 1995. How scientists really reason: scientific reasoning in real-world laboratories. In: Sternberg RJ and Davidson JE (Eds). The nature of insight. Cambridge, MA: MIT Press.

Ellison AM. 2010. Repeatability and transparency in ecological research. *Ecology* **91**: 2536–39.

Ernest SKM, Valone TJ, and Brown JH. 2009. Long-term monitoring and experimental manipulation of a Chihuahuan Desert ecosystem near Portal, Arizona, USA. *Ecology* **90**: 1708.

Hamilton MP, Graham EA, Rundel PW, *et al.* 2007. New approaches in embedded networked sensing for terrestrial ecological observatories. *Environ Eng Sci* **24**: 192–204.

Hampton SE and Parker JN. 2011. Collaboration and productivity in scientific synthesis. *BioScience* **61**: 900–10.

Heidorn PB. 2008. Shedding light on the dark data in the long tail of science. *Libr Trends* **57**: 280–99.

Jones MB, Schildhauer MP, Reichman OJ, and Bowers S. 2006. The new bioinformatics: integrating ecological data from the gene to the biosphere. *Annu Rev Ecol Evol S* **37**: 519–44.

Keller M, Schimel DS, Hargrove WW, and Hoffman FM. 2008. A continental strategy for the National Ecological Observatory Network. *Front Ecol Environ* **6**: 282–84.

Kelling S, Hochachka W, Fink D, *et al.* 2009. Data-intensive science: a new paradigm for biodiversity studies. *BioScience* **59**: 613–20.

Kwa C. 1987. Representations of nature mediating between ecology and science policy: the case of the International Biological Programme. *Soc Stud Sci* **17**: 413–42.

Michener WK, Beach JH, Jones MB, *et al.* 2007. A knowledge environment for the biodiversity and ecological sciences. *J Intell Inf Syst* **29**: 111–26.

Michener W, Vieglais D, Vision T, *et al.* 2011. DataONE: Data Observation Network for Earth – preserving data and enabling innovation in the biological and environmental sciences. *D-Lib Magazine* **17**: 1–2.

NABCI (North American Bird Conservation Initiative). 2011. The state of the birds 2011 report on public lands and waters. Washington, DC: US Department of the Interior. www.stateofthebirds.org. Viewed 19 Sep 2012.

Noor MAF, Zimmerman KJ, and Teeter KC. 2006. Data sharing: how much doesn't get submitted to GenBank? *PLoS Biol* **4**: e228.

Palmer MA, Bernhardt ES, Chornesky EA, *et al.* 2005. Ecological science and sustainability for the 21st century. *Front Ecol Environ* **3**: 4–11.

Piwowar HA, Day RS, and Fridsma DB. 2007. Sharing detailed research data is associated with increased citation rate. *PLoS ONE* **2**: e308.

Price DJ. 1963. Little science, big science. New York, NY: Columbia University Press.

Reichman OJ, Jones MB, and Schildhauer MP. 2011. Challenges and opportunities of open data in ecology. *Science* **331**: 703–05.

Spengler SJ. 2000. Bioinformatics in the Information Age. *Science* **287**: 1221–23.

Stokstad E. 2011. Open-source ecology takes root across the world. *Science* **334**: 308–09.

Strasser BJ. 2008. GenBank: natural history in the 21st century? *Science* **322**: 537–38.

Weinberg AM. 1961. Impact of large-scale science on the United States. *Science* **134**: 161–64.

Zimmerman AS. 2003. Data sharing and secondary use of scientific data: experiences of ecologists (PhD dissertation). Ann Arbor, MI: University of Michigan.

