Incentives for Data Sharing in
Ecology, Evolution, and Organismal Biology
February 19-20, 2009
Workshop Report

BACKGROUND

A Joint Working Group on Data Sharing and Archiving (JWG), representing major professional societies that publish ecology, evolution, and organismal biology journals, was formed at a September 2004, NSF-sponsored workshop on data sharing and archiving, hosted by the Ecological Society of America (ESA). Attendees adopted a consensus statement that "Our vision as members of the scientific community is to promote the advancement of science through the process of documenting, archiving, and making available the research information and supporting data of published studies." In support of that vision, the JWG made the following recommendations:

- Facilitate continuing communication among professional societies on data sharing and archiving issues via a dedicated web site and periodic e-mails;
- Widen participation in these activities by professional societies and international organizations; and
- Support three workshops to (1) develop a strategy for creating data registries that describe datasets and provide information on how to access them, (2) identify, and develop means to reduce or eliminate, cultural and other barriers to data sharing, and (3) develop a set of requirements and recommendations for data centers in ecology, evolution, and organismal biology.

The first of these three workshops, "Data Registries for Ecology, Evolution, and Organismal Biology," was held July 11-12, 2006, in Washington, DC. Twenty-five participants representing 16 professional societies and nine other organizations assembled to work toward three goals:

- Identify a set of common needs for, and desirable features of, data registries for ecology, evolutionary biology, and organismal biology, based on an understanding of existing resources.
- Develop recommendations, as appropriate, for shared or independent data registries for the disciplines and societies represented.
- Develop preliminary plans for implementing those recommendations.

The second workshop, "Data Centers for Ecology, Evolution, and Organismal Biology," was held December 8-9, 2006, in Santa Barbara, CA. Thirty-two participants representing 14 professional societies and 11 other organizations assembled to work toward three goals:

- Identify gaps between existing data centers and needs, including specific issues such as quality assurance procedures needed for contributions to centers, types of data that should be archived, etc.
- Identify roles of professional societies, funders of research, and users of research in developing – or encouraging the development of – data centers, along with where data centers should be housed and who should operate and maintain them.
- Assess likely cost to establish and maintain data centers required to meet community needs, including identification of potential funding mechanisms and models for data centers.

The third workshop, "Obstacles to Data Sharing in Ecology, Evolution, and Organismal Biology," was held May 30-31, 2007, in Durham, NC. Thirty-nine participants representing nine professional societies and 22 other organizations assembled to work toward two goals:

- Clearly delineate what barriers exist to data sharing, for example, intellectual property concerns, proprietary and confidential business information, handling of sensitive data such as locations of

endangered species, lack of training in data sharing software, national or economic security concerns, etc.
- Develop recommendations to reduce or eliminate those barriers, for example, publication policies that encourage or require data sharing, means of providing at least limited access to business or sensitive data, and development of easily accessible training programs.

A fourth workshop, "Developing Incentives for Data Sharing in Ecology, Evolution, and Organismal Biology", was held February 19-20, 2009, in Washington, DC. 24 participants representing researchers, publishers, and funders assembled to work toward two goals:

- Identify incentives and recommend steps to overcome barriers to productive sharing of scientific information from the perspective of funders, researchers, and publishers.
- Develop ideas for products that will help implement recommendations from the Data Sharing Workshop series.

This report summarizes the discussions and recommendations from the fourth workshop. Reports on the first three workshops are available at http://www.esa.org/science_resources/datasharing.php.

INTRODUCTION AND WELCOME

Cliff Duke (Ecological Society of America) welcomed workshop participants and provided background information on the Data Sharing Initiative. He charged participants to discuss current data sharing policies and incentives, including incentives by funding organizations and government to support sharing of data.

Jim Reichman (past Director of National Center for Ecological Analysis and Synthesis [NCEAS]) gave a talk, "Why share data (what is the value to researchers of sharing data?)". There are many reasons for researchers to share data in order to help the scientific community continue to move research forward. There is also an argument that, if data collection is publicly funded, then there is a responsibility on the researcher to make that data publicly available. A study conducted by NCEAS revealed that, in return for sharing data, data providers desired formal acknowledgement and financial assistance to cover the time it takes to format data for public use. The survey also showed that people were willing to satisfy *more* severe conditions to gain access to other data than they would themselves require. However, obstacles such as settling issues of intellectual property rights, particularly when data have been funded by multiple sources and potentially from different countries, have not been sorted out yet.

Reichman then discussed issues related to open access publishing. Open access can have impacts on authors, readers, and publishers. Authors could see citation rates increased; readers can have increased access to sources and links in papers; and publishers could need changes to their income and profit structures. Several models exist for open access publication, including having authors pay out of grants for open access, incorporating "delayed" open access so articles would be accessible for a cost before they would be free, and using a "threshold" model so institutions would have open access to research until they reached a certain level of readership.

INFORMATIONAL PRESENTATIONS ON EXISTING DATA SHARING POLICIES

Michael Whitlock (University of British Columbia and former editor of *The American Naturalist*) discussed data sharing in evolution journals. It is important for researchers to share their data in order to provide an avenue for error checking, to allow new methods for meta-analysis and new interpretation of existing data, and to increase citations. However, most ecology and evolution journals can not require

data archiving because the data repositories do not exist yet. There is an initiative to create a data archive, Dryad (www.datadryad.org), which is currently accepting data by email while it is in beta version.

Five major evolution journals adopted a common data sharing policy at the same time so no journal would suffer from being the sole source to encourage data sharing. The draft policy states, "The [journal] requires, as a condition for publication, that data used in the paper should be archived in an appropriate public archive, such as GenBank, Treebase, or Dryad. The data should be given with sufficient details that, together with the contents of the paper, allow each result in the published paper to be re-created." This policy provides for three accommodations: 1) data can be archived with a one-year embargo on public access if desired by the author(s); longer embargos can be considered upon application to the journal editor; 2) archived data should be cited fairly, and journals should encourage citation of the original paper, not accession numbers; and 3) archiving is required only for data used in the paper being published.

William Michener (University of New Mexico and Director of the LTER Network Office) discussed the history of data management policies for the Long Term Ecological Research Network. While certain polices have been in effect since 1990, a new set of requirements released in 2005 gave “teeth” to the policy. Components include making data available even if the principal investigator (PI) leaves; dividing data into two types so that access to sensitive data (e.g.: locations of endangered species, human data) can be protected; and encouraging the use of data repositories rather than individual websites for data archiving. Michener also discussed polices from other groups. For example, the Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO) requires that data be available within one year of collection and within two years to the public via data catalog.

Michener’s final points encouraged the scientific community to look to universities and libraries to help with data sharing. We need a wide range of partnering organizations, and we need to re-envision academic cyberinfrastructure (CI). We also need to start training students in bioinformatics, similar to how all ecology students are trained in statistics. One final note was that it will not be possible to store all data. In the year 2007, more data were collected that can ever be stored so we will need processes for deciding what should and should not be retained in the long term.

Charlotte Gray Hudson (Pew Charitable Trusts) and Chris Mentzel (Moore Foundation) presented their perspectives as funders on data sharing initiatives. Pew has not incorporated any data sharing language into its grant process yet; however, it has been built into deliverables. They see data sharing manifesting more in websites rather that in data archives. A majority of their work is data synthesis and not data collection; however they are recognizing that these results could still be shared. Hudson posed the question: If data sharing by researchers was required by funders, would the funding be worth the additional time and effort, or would researchers not want the grant?

The Moore Foundation created its first policy on open access to data in 2005. However, they found that the 9-page set of guidelines was not being adhered to, so now they have switched to a “data sharing philosophy”. This includes an expectation that every grant will have a data sharing plan that describes the data, management of the data, and how they will be shared. The Foundation is interested in considering incentives not related to funding, and would like to move the data sharing initiative beyond ecology to science in general. Data curation is also important because the large deluge of data will be unmanageable if it is not filtered.

Chris Greer (National Science and Technology Council) is part of a diverse Interagency Working Group on Digital Data that has been charged with developing and implementing a strategic plan to assure that governmental scientific data are useable and accessible (see http://www.nitrd.gov/about/Harnessing_Power.aspx). This includes a broad interpretation of data, encompassing videos, models, and simulations. The report was released on February 19, 2009, and include guiding principles, such as: 1) preservation is both a government and private sector responsibility; government should be a leader and a partner in these efforts; 2) “communities of practice”

are essential – data for different communities will be different and so there will not be a "one-size-fits-all" solution; 3) not all data need to be preserved indefinitely; both data managers and scientists need to be involved in these decisions.

The report also provided these recommendations: 1) organizations should be responsible for managing data; 2) departments and agencies should develop their own data policies; and 3) proposals/projects that will generate data should have a data management plan from the outset. The working group has already begun developing the first two recommendations by looking at the best practices that exist so far and determining which principles should be covered in a policy and considering how each agency could refine or make its own policy.

Tom Moritz (Internet Archive) discussed the "commons" and the importance for scientific knowledge to be available to the public. He suggested that this knowledge might exist on an "Ethics Spectrum", where information pertaining to human health, conservation, and agriculture would be as open to the public as possible, but that there would be ethical reasons to restrict access to information for other topics, such as nuclear technology. Since public libraries have historically been the "protected areas" of the knowledge commons, libraries could continue this role in managing open-access scientific data.

In particular, Moritz talked about the "Conservation Commons", which "promotes and enables conscious, effective and equitable sharing of knowledge resources to advance conservation". The Conservation Commons promotes free and open access to data, information and knowledge for conservation purposes, provides a mutual benefit by encouraging users both to contribute to and to use the data, and has a disclaimer of rights and responsibilities, so that authors maintain full right to attribution when their data or knowledge are used. About 65 organizations have formally endorsed these principles, including The Nature Conservancy, IUCN – The World Conservation Union, Smithsonian Institute, US NASA, and World Wildlife Fund International.

Dave Schindel (Smithsonian Institution and Executive Secretary for the Consortium for the Barcode of Life) discussed the Barcode of Life Initiative, which was started in 2003 with a single publication. This project is an example of data sharing within one community focused on taxonomy. Instead of using characters and concepts to group specimens of the same species, scientists can assign a "DNA barcode" – a short gene sequence taken from standardized portions of the genome that is unique to the species. Each animal then has a barcode, equivalent to each item in a grocery store having a unique barcode. In 2004, the Consortium for the Barcode of Life was created and since then more than 500,000 records for 50,000 species have been gathered. Most of these data are password-protected under ownership of authors and not in GenBank. With GenBank, a data standard was created and a species can be entered by filling out 10-15 fields. Schindel emphasized that this system has worked for this community. The Public Library of Science has begun to publish fast-track, short, standardized format papers in PLoS ONE to release the information on the barcode. There is also a long term effort devoted to data curation of the barcode records by using the public commenting function in PLoS ONE to point out an error in the record. This serves as a curatorial feedback mechanism.

Schindel also promoted an upcoming conference, e-Biosphere, taking place in London in early June, 2009. The Conference will pose two challenges to the community: 1) how much faster can taxonomists describe the millions of undiscovered species?; and 2) what will users of biodiversity information need in the coming years? What new databases and capabilities will they need?

BREAKOUT DISCUSSIONS

Following the initial plenary presentations, participants moved into two breakout groups, one of research funders and one of researchers.

Funders Group

The Funders breakout group based their discussion around four questions: 1) What might your organization be willing to do to encourage data sharing?; 2) How might your organization's goals be enhanced by incentives for data sharing?; 3) How are your organization's goals enhanced by making data sharing policy mandates?; and 4) What are you doing, or not doing but could be to encourage data sharing by grantees?

Many organizations are starting to encourage data sharing. The USDA grants program has yearly PI meetings to discuss data; the National Phenology Network is linking ecological data sets; the National Ocean Service is starting to build databases and track metadata tools. However, obstacles still exist. For example, the DoD-Legacy program faces dilemmas with data that are protected under military rules. Some people are concerned about security; some data should not be readily available to the public, such as locations of populations of endangered species. From a foundation perspective, many foundations have several different fields and types of research. Not all grants produce data, so each grant needs its own plan.

There are still many technological obstacles with data sharing. Projects that are initiated with data sharing in mind are easier since the data collection is designed with collaboration in mind. It is much more difficult to shift data from closed to open access. One incentive would be to create hosted analytical space from the beginning of projects. The data would be archived, backed up, and accessible from various points.

Other incentives include data backup, QA/QC, standardization, and publication acknowledgements. We need to make it easy for researchers to add to their metadata. This could be done within libraries. The people in charge of cataloging data are those who can create useful metadata, not necessarily the PIs. Perhaps an incentives program would provide a person or employee of the agency that would go to individual PIs and help them archive existing data sets and devise plans for future data.

Funders and publishers are instituting policies that will require first and foremost, the creation and registering of metadata for all publications (in some cases datasets are recognized as publications), and simultaneous or delayed publication of datasets as a requirement for publication or completion of grant requirements. NSF representatives noted that their organization has no single policy in regard to data sharing and access, but that several boards have established policies in this regard. The LTER protocols were cited as a successful model that has been effective in encouraging collaboration and in stimulating new and integrated science products.

Researchers Group

The Researchers group focused on three discussion points: 1) create recommendations with respect to research, development in education and outreach, and cyberinfrastructure (CI); 2) create an agenda for funders; and 3) discuss the role of data centers and future actions regarding data sharing.

1. Suggestions for R&D:

- Community data sharing has issues with the tragedy of the commons. The community should be discussing the opportunity costs of NOT sharing data.
- Encourage long-term commitment to sustaining data centers with a "continuity budget."
- Monitor the use of data centers – how much people are uploading and using the center. Encourage adaptive management and keep up on the technology for existing centers.
- Develop tools for automated harvesting of metadata.
- Rather than start new programs, stitch together programs that already exist.
- Make sure that the true life cycle cost of projects includes time and funds for data sharing.

- Incorporate data sharing into proposal reviews and budgets
- Develop informatics curricula to teach best management practices for people's own data. Short modules could be incorporated into biostatistics courses.
- Link data centers so they are all accessible from each other.

2. Agenda for Funders

- Establish common standards, e.g.: metadata registries and open-source repositories.
- Funders could hold back final payment of grant until data have been made available.
- Need to factor in additional costs for researchers to prepare data for sharing.
- Support undergrad and graduate resources, training, etc. on data sharing.

3. Role of Data Centers

It is important that researchers read the original papers that first analyzed a given set of data. A proposal for a "three-tiered" approach to data sharing was discussed. The first tier included data that were shared in a "do-it-yourself" manner and maintained on personal websites. The second tier is "self-curated," by following some standards, and the third tier would be a peer-reviewed data paper. This third tier would be valuable to the community. Community interaction at the second tier could move a dataset into the third tier.

The Researchers group ended with the question: How will data sharing policies address derived data sets? If a researcher cleans up and changes a data set that was not originally his/hers, how will that be referenced and attributed?

The Researchers group also noted that institutional barriers to release of data are still significant, although changing. Impediments range from university policies and concerns about copyright law, to government review processes that are out of sync with the rapid pace of moving data from the laboratory or field to the web. Also, mechanisms for attribution of data sharing and creation of database and database tools are either non-existent or do not carry the same weight in academic and government research that peer review journal publications do in the evaluation of individual researchers work. In the view of the participants, this is changing, and much of the difference is generational. Younger researchers are more "web-savvy" and more willing to recognize the value of data publishing in moving a discipline forward. Both groups also recognized that there is a difference between data archiving and data sharing and that this distinction may define the level of participation in or support for each activity among researchers – data archiving is not necessarily synonymous with data sharing.

PLENARY DISCUSSIONS

In a plenary discussion after the presentations and breakout group meetings, attendees were asked to consider incentives for both individuals and groups/organizations, as well as any ethical/social justice components of data sharing.

Brian Wee provided background on the National Ecological Observatory Network (NEON). Planning for NEON began in late 2004 with the idea of large-scale data collection to monitor effects of changing climate and invasive species. A network of 60 sites across the United States would collect data on 100-120 variables to quantify the state of the nation's ecosystems. NEON is in the process of developing data sharing policies. There is a question of data ownership. If a researcher collects data with NEON, who owns the data? NEON will need agreements on this with its funding institutions.

Another suggestion was that high-level discussions should be held with the presidents and research managers at major universities. Do they already have policies and procedures in place for university-owned data?

If major universities are hubs for data, how does this affect community colleges, tribal schools, etc., that are excluded from technology and access? Is there an obligation to supply data to these groups? No NEON sites are on tribal lands, although one is nearby.

Everyone agrees that new programs, such as NEON, need to examine existing policies and try to take advantage of what already exists. Wee explained that NEON is adopting existing standards, such as those from government agencies.

It is also vital to consider the sustainability of programs since this is important for long-term preservation. If NEON collects data and then loses its funding, will that data be lost?

WORKSHOP CONCLUSIONS AND RECOMMENDATIONS

- Focus on metadata. Quality metadata is the gateway to data and information exchange; until standards are achieved within a community of practice, the productivity and value of information and especially data are limited.

- The first level of cooperation for facilitating data sharing is a community of practice. As amorphous as the term seems, communities are fairly effective in self-identifying, as has been shown by the progress associated with data sharing thus far. Success stories include NCEAS and NESCent in the ecological and evolutionary biology communities, respectively. The lessons from open access software in the computer industry are indicative that data sharing evolves to build and share information collection, management, exchange and other software tools. GenBank, which is supported by the National Institutes of Health, Morphbank, supported by NSF, and FishBase are examples of robust data sharing efforts that have become standards in their communities of practice and are increasingly self-sustaining.

- Throughout this and previous workshops on data sharing, participants have agreed that metadata creation and dataset archiving and sharing should be a simple and quick task in order to encourage and facilitate participation in these activities. Simple, free and/or cheap, and widely available software tools for creating and registering basic standard metadata records (either derived from publications or cataloged manually), and for enabling digital dataset submission, may play a significant role in increasing and improving within- and cross-disciplinary data archiving and sharing.

- A challenge for ecology and other interdisciplinary fields is that these communities require cooperation and connections across communities of practice. Establishing standards and creating data sharing connections will require institutions – agencies, professional societies, governments, international bodies such as treaty organizations and foundations – to create incentives for information sharing and access.

DISCUSSION OF NEXT STEPS AND ELEMENTS OF WORKSHOP REPORT

- ESA staff will prepare a draft Workshop Report, circulate it to participants for input, and then distribute a final version to be shared with society leaders.

- ESA will make the workshop report available to NSF as part of its responsibilities under the workshop grant.

- Have a core set of graphs and slides created about the data sharing workshops and circulate these among all participants. As participants go to conferences, they can add their own information to individualize their presentations.

- Other suggestions included hosting another activity at the ESA Annual Meeting (perhaps geared towards students and data management skills), adding data sharing resources to ESA's website, and writing a Science Policy Forum with a clear set of recommendations from the workshops.

- One or more summary papers from all of the data sharing workshops will be prepared for publication in *Bioscience* or *Frontiers in Ecology and the Environment*. Developing an *Issues in Ecology* on data sharing will be considered.

- Representatives from the professional societies, journals, and networks will continue to work toward shared policies regarding publication and attribution for data. Academic institutions and government agencies will need to update polices regarding the value of publication of data and database tools and techniques to provide increased incentive for their scientists to publish data in a timely fashion.

- David Schindel, with others, will attend the upcoming International Conference on Biodiversity Informatics, "Biosphere 09", June 1-3, 2009, hosted by the Natural History Museum, London, GB. This conference will be a showplace for the latest breakthroughs in information science and the biological and environmental sciences. Additional information at www.e-biosphere09.org.

Incentives for Data Sharing in Ecology, Evolution, and Organismal Biology
February 19-20, 2009
Washington, D.C.

ATTENDEE LIST

Attendee	Affiliation
Mike Bowers	Cooperative State Research, Education and Extension Service
Susan Cameron	Harvard University/Student Section, Ecological Society of America
Chris Greer	National Science and Technology Council
Clifford Duke	Ecological Society of America
P. Bryan Heidorn	National Science Foundation
Charlotte Gray Hudson	Pew Charitable Trusts
Wayne Litaker	National Oceanic and Atmospheric Administration
Martha Maiden	National Aeronautics and Space Administration
Corrie Mauldin	Ecological Society of America
Peter McCartney	National Science Foundation
Chris Mentzel	Gordon and Betty Moore Foundation
William Michener	University of New Mexico
Pedro Morales	Legacy Program, Department of Defense
Thomas Moritz	Internet Archive
Rachel Muir	United States Geological Survey
Greg Reams	United States Forest Service
Rhett Rebold	DISDI Program, Department of Defense
Jim Reichman	self
David Schindel	Consortium for the Barcode of Life
Elizabeth Sellers	National Biological Information Infrastructure
Brian Wee	National Ecological Observatory Network
Michael Whitlock	American Society of Naturalists
Aleta Wiley	Ecological Society of America
Bruce Wilson	Oak Ridge National Laboratory