Society for Conservation Biology
A global community of conservation professionals

January 12, 2012

Ted Wackler
Deputy Chief of Staff
Office of Science and Technology Policy
Attn: Open Government
725 17th Street, NW.
Washington, DC 20502

Submitted via e-mail to digitaldata@ostp.gov

Dear Mr. Wackler,

The Society for Conservation Biology (SCB), a global community of conservation professionals which publishes Conservation Biology, among other journals, submits these comments in response to the request by the Office of Science and Technology Policy (OSTP) for input on the Administration's interest in enhancing public access to digital data generated in federally funded research. In the following comments, we borrow substantially from a draft prepared by our sister societies in the Ornithological Council, a consortium of twelve scientific ornithological societies in the Western Hemisphere and from comments submitted individually by our President, Paul Beier.

Conservation Biology is rich in data that are underutilized because they are not accessible. Decades of data are disappearing rapidly and irretrievably because the scientists who collected the data had no opportunity to archive it in a physical or electronic form. Whether on paper or in some kind of electronic medium, datasets collected over the past century could contribute greatly to our knowledge of conservation biology.

Our organization strongly supports the concept of archiving and sharing these data. Sister societies have investigated and discussed the possibility of developing an archive for the types of data generated in different forms of biological research but found that the cost is prohibitive and that it might not be realistic to expect that scientists will voluntarily undertake the somewhat burdensome effort of learning metadata standards and routinely labeling their data for deposit into an archive.

As a preliminary and key issue, we stress the need to allow researchers to have exclusive access to and use of their data for a time period sufficient to allow them to complete their publications. This time period must be flexible; in our field, long-term studies can stretch over decades. The "reward system" for scientists in both academia and in federal agencies stresses publications. The number and quality of publications is a large factor in determining promotion and tenure and strongly affects the researcher's success in

obtaining grant funding. We assume that OSTP is fully aware of the fact that the misappropriation of a researcher's data could have substantial negative impacts on the researcher's career and will take care to assure that any public access policy includes ample protections for the researcher.

As a second key issue, we would like to address something that seems to be outside the scope of the OSTP request and existing agency data management requirements, probably because it would be impossible to impose these requirements retroactively. We would like to stress that if resources are available, the government should commit those resources to help "stabilize" those data, convert them to a digital format, and submit them to appropriate data repositories. The data collected a decade ago or a century ago are, in our field, at least as valuable as the data collected today, if not more so, as these baselines are necessary to assess change. The attics full of paper, note cards, field notes; the offices full of punch cards, floppy disks, and magnetic tape – all need proper storage to guard against physical loss and all should be digitized and contributed to publicly accessible repositories. We cite the example of the North American Bird Phenology Program created by the Patuxent Wildlife Research Refuge of the U.S. Geological Survey. Using volunteers and a high-speed scanner, this remarkable program preserved six million hand-written note cards recording bird migration observations, dating back to 1881. The scanned records were then uploaded to the internet to make it possible for volunteers to enter the data into a database. The USGS and the other partners of the National Phenology Network provide analytical tools, guidance documents, and other resources. More recently, the U.S. Bird Banding Lab was able to stabilize decades of hand-written records by scanning and it is hoped that funds will be made available to make these critical data available to researchers by digitizing the data and making them available on a public access website. To date, researchers and others have been able to access these data only by making a request to Banding Lab staff who would then retrieve the physical records for copying and mailing. The records were at extreme risk of physical deterioration or loss, having been stored in a variety of facilities that were subject to rodent infestation, fire, dampness, and flooding.

Therefore, we strongly encourage OSTP to work with the Office of Management and Budget and Congress, as appropriate, to provide funding and direction to the agencies to stabilize existing physical data records, to digitize those records, and make them available on publicly accessible databases. These processes should not be limited to agency-held data but should be opened to private researchers as well.

We would also like to address certain of the questions asked by OSTP, as follows:

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Response: The key issue here is funding. Developing and maintaining these systems is costly. The intricacy involved in creating any one metadata standard is substantial. Interoperability is a daunting challenge. In our discipline, for instance, DataOne <www.dataone.org> is intended to "ensure the preservation and access to multi-scale, multi-discipline, and multi-national science data. DataONE will transcend domain boundaries and make biological data available from the genome to the ecosystem; make environmental data available from atmospheric, ecological, hydrological, and oceanographic sources; provide secure and long-term preservation and access; and engage scientists, land-managers, policy makers, students, educators, and the public through logical access and intuitive visualizations." The five-year NSF grant alone amounts to $15,257,190 from the Office of Cyber Infrastructure and it is supplemented by support from the NSF Computer and Information Science and Engineering Directorate (CISE) Pathways Computational Sustainability, the NSF INTEROP Programs, NASA, the Leon Levy Foundation, the Moore Foundation and (until its recent demise), the National Biological Information Infrastructure of the U.S. Geological Survey.

The complexity of these systems requires that they be done right; if not, the end result is a system that hampers, rather than facilitates public access. The federal government must be willing to commit the resources to enable excellence or the undertaking is not worthwhile. We would have an expensive warehouse where nothing can be found, much less retrieved.

We would draw your attention to an article addressing these issues in Science Magazine. The citation, abstract and some of the recommendations follow:

11 FEBRUARY 2011 VOL 331 SCIENCE www.sciencemag.org

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.

Some fields such as 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.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Assure that researchers have the ability to control the release date. Do not require release until the researcher has had adequate time to publish the research utilizing a given data set. Recognize that in some fields, research may extend over decades. For instance, studies of long-lived organisms will typically continue over the full life-cycle of the organism. A researcher will likely publish papers throughout this period, but later papers will often make use of data collected at a much earlier stage of the study. Consult with scientific societies to determine the appropriate maximum duration for the sequestration of a given data set.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Consult with the professional societies. We can provide the data and insight as to the policies and practices that will make it possible for our members to archive and share data without jeopardizing their intellectual property interests. We can also provide information about the ability of our discipline to create and maintain these repositories and the appropriate metadata standards. We can identify gaps in opportunities for data management. In ornithology, the existing repositories, though stellar, simply cannot accommodate many kinds of data collected by ornithologists. We have, as a result of the NSF data management plan, been collecting information about all potential data repositories that may be suitable for this kind of data, and we are still finding significant gaps. At the moment, NSF's data management website simply directs those who are unable to find an appropriate public repository to "Contact the cognizant NSF Program Officer for assistance in this situation." We suspect that if NSF were to attempt to compile a comprehensive list of relevant data repositories, these gaps would be quite evident.

We can also compile and provide data about the range and median grant size in our discipline. This information should be taken into account before imposing another time-consuming grant requirement on researchers. The OSTP notice mentions that the NIH requirement applies only to grants with direct costs exceeding $500,000 in a single year. In our discipline, that threshold would exclude most grants. For instance, the average grant size made by the NSF BIO program in 2011 was $149,238. In 2010, it was $140,064 <http://dellweb.bfa.nsf.gov/awdfr3/default.asp>. Most NSF grants in our discipline come from the Division of Environmental Biology (DEB) or the Division of Integrative and Organismal Systems (IOS). In DEB, the average grant in 2010 was $95,649 and in 2011, it had declined to $85,919. In IOS, the average grant size was $150,000 in 2010 and $$151,181 in 2011. Smaller grants simply do not allow the researcher to hire administrative staffers or other technicians to handle this additional work.

If no additional funding is provided, the data management requirements could constitute an unfunded mandate such as would trigger the provisions of 2 U.S.C. §1501. We recognize that the Administrative Procedure Act exempts matters "relating to agency management or personnel or to public property, loans, grants, benefits or contracts" and that therefore, a formal rulemaking as would trigger the Unfunded Mandates Reform Act (UMRA) would likely not occur. Nonetheless, the agencies have made it a practice to use notice-and-comment procedures outside the Federal Register process for this and other policy matters. These quasi-rulemakings should be regarded, for the purpose of the required UMRA analyses, as the equivalent of a rulemaking. Therefore, any agency that wishes to mandate data management should be required to conduct an "UMRA-like" analysis to assure that the requirements are the least costly, least burdensome, or most cost-effective option that achieves the objectives of the rule, or explain why the agency did not make such a choice (2 U.S.C. §1535).

The scientific community should also be consulted with regard to the release of certain types of data. For instance, we have long been concerned about the potential online, public access release of location information associated with bird banding. Some of the birds banded are, of course, legally protected at the federal or state level. Information about the location of banding could facilitate activity that is prohibited under the Endangered Species Act. Other species, protected only under the less comprehensive prohibitions of the Migratory Bird Treaty Act, are very vulnerable to disturbance during the breeding period. If the public could use the location data associated with bird banding to determine breeding locations, the disturbance resulting from human presence could lead to failed breeding attempts. This outcome would contradict Executive Order 13186.

As noted, some data could be used by unscrupulous persons to kill, capture, or harm individual animals or plants. Many agencies (Arizona's Heritage Database Management System, for instance, and other state databases) have largely solved this problem, however, using two simple measures: (1) The publicly available data consist only of low-resolution maps with locations "fuzzed" by up to a few km. This provides enough preliminary information for a potential user to determine if the data cover the area of interest to the user. (2) Precise location data are provided only to legitimate requestors

who agree to specific terms on use of the data, including agreements not to depict or share precise locations in any way.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

For species occurrence data, the costs are miniscule and the benefits are large. We suggest that OSTP might for now require data sharing only for similar types of low-cost high-benefit data. OSTP and other agencies could use the experience to start to produce reliable estimates of long term costs and benefits that could be used to guide future decisions.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

The Society for Conservation Biology publishes several scientific publications. SCB could work with our publisher to require authors to archive their species location data with appropriately coordinated repositories. However, if only SCB took this step, some authors would submit elsewhere to avoid this extra responsibility. But a broad consortium of professional societies in ecology (SCB, Ecological Society of America, The Wildlife Society) and a handful of dominant publishers (e.g., Wiley-Blackwell, Elsevier, Springer-Verlag) could create a new culture in which data-sharing is viewed as a responsibility of publishing. Our President has appointed a Task Force in SCB to investigate how SCB could start a dialogue with our sister professional societies and the publishers of their journals to start to create this culture. It will take years, and there will be strong resistance from some academic PIs, but this is an achievable long-term goal. Again, it makes sense to start with low-hanging fruit (e.g., species occurrence data); once the new culture of sharing has been in place for a few years, I think it will become obvious which other types of data to share, and how to share them.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

As noted above, grants in our field typically do not permit researchers to hire staff to undertake the work associated with effective metadata labeling and deposit of data. There is no point in warehousing data if it is not done in such a way as to make the data easily retrievable and to assure that subsequent users are able to identify the characteristics of those data so they can determine if they are appropriate for the later use. Without additional funding, data repositories are not likely to be of adequate quality and any resources devoted to them will have been wasted.

This is not a hypothetical concern. The U.S. Geological Survey devoted more than a decade of effort to develop the National Biological Information Infrastructure. It is now

being dismantled; it never began to approach the original goal of providing access to distributed data.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

For some types of data, ensuring compliance will be difficult, but it should be relatively easy for species occurrence data. Federal funders of biodiversity-related research (NSF, USDA, DOD SERDP, EPA) could require the Data Management Plan in each proposal to list the species for which occurrence data will be collected. Funders should convey this information to a repository that is well integrated with others, which would need staff persons to track compliance and report non-compliance to all federal funders.

One more drastic measure is worthy of consideration: The OSTP and OMB could set out procedures for identifying institutions with a pattern of non-compliant PIs and baring such institutions from future federal grants and contracts for a period of time. This would motivate universities and other research institutions to monitor compliance.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

By citing in their proposed and final rulemakings more thoroughly the peer reviewed journals and the data reported and analyzed therein, and by working with Congress to help their committees and the Congressional Research Service to do the same in legislative and investigative and oversight committee reports.

Also by making information that will be available publicly someday available sooner in some cases. For example, in addition to considering Federal purchasing of rights to copyrighted material, OSTP might consider working with expert Federal agencies and the Federal Office of Trademark, Copyright and Patents to determine the extent to which currently patented procedures and devices that could help solve serious societal problems, such as increasing energy efficiency and reducing pollution, or sequestering carbon with bio-char produced in biologically sound and safe ways, are being fully deployed and if not, what level of payment would be appropriate for an eminent domain-style assumption of part of or the remaining years of that patent by the Federal Government. Agencies could review indexes of patents or other descriptions of them with the help of the Patent Office. They could then ask scientific societies to help them evaluate those that might be more useful if provided to the public at an earlier point.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

For a number of years, we have discussed this very question with regard to the potential release of bird banding data. It has been the practice of the Banding Lab to interact with

those who request data and to remind them of the professional standards for attribution and credit. This interaction is possible only because data requests are made by individual contact to a staffer who then transmits the data to the requester. In fact, the Banding Lab website makes no mention of these professional standards. The U.S. Bird Banding Lab Advisory committee could not devise a more robust solution, saying that a web-based public access site should be developed and that In consultation with banders and users of banding data, review and revise the current policy for use of banding data, and require all data users to agree to this policy. The BBL should also encourage the adoption of this policy by ornithological societies and scientific journals as part of their scientific code of ethics."

The reality is that there is no effective mechanism to force users to give appropriate attribution and credit. It may be evident, given the age of the data or the geographical or temporal range of the data that the author did not collect all the data used in the paper. In those cases, editors will likely insist that the author provide attributions. However, there will be many cases where this is no evidence that the data used were collected by other than the author, and in those cases, there is really no adequate solution.

Therefore, the only means to protect a researcher who is still publishing papers based on a given dataset is to allow the researcher to determine the date of release of the data to the public, as described above, subject to standards that are appropriate to that particular discipline.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data?

Our task force may be able to help with this soon but we have no comment on this yet.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

In the taxonomic sciences, extensive effort has gone into the development of a metadata standard known as the Darwin Core. Numerous extensions have been developed that will support the addition of "ancillary" data such as ecological conditions, and weather data. We hope that there will someday be extensions for the behavioral data that is commonly collected in biological and related research.

The use of this common metadata standard and extensions would permit interoperability with any other system that uses the same standards. For instance, the Darwin Core has led to the development of ORNIS, HerpNet, MANIS, and FishNET (birds, herps, mammals, and fishes) and these are integrated with GEOLocate, AmphibiaWeb, Map of Life, Specify, Arctos, DataONE, Encyclopedia of Life, and Animal Diversity Web.

These repositories and the metadata standards were initiated by the community and achieved with federal funding. Other organizations (most also federally funded) then built user tools and applications, such as the Avian Knowledge Network at the Cornell Lab of Ornithology. This project also received significant federal funding.

However, no amount of scientific zeal and energy can achieve this kind of result without significant federal funding. Unless the federal government is willing to continue to devote appreciable sums, the government and the public cannot expect to achieve the goal of providing public assess to data derived from federally funded research.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Response: Science knows no geopolitical boundaries. Scientists have long been working on an international basis to develop metadata standards. The Global Biodiversity Information Facility, established in 2001, already holds 8,594 datasets to which access is free and unrestricted. However, the sole U.S. representative to GBIF is a single employee of the now-terminated National Biological Information Infrastructure. The NBII termination page states with regard to GBIF that "While USGS does anticipate continued collaboration with some of these activities, we have yet to determine at what level this will occur." We are informed that it is likely that USGS will continue to participate at the minimal level (i.e., one FTE) that was the case prior to the termination of the NBII.

The federal agencies must commit to increased participation in these international bodies, and commit the necessary resources for that participation.

If the federal government is unable or unwilling to continue funding this activity at an adequate level, then it should hold in abeyance all but the most compelling and reasonable mandates that scientists submit data to any repository. If there is no assurance that the repositories will persist and will be properly managed, and that there will be a continued development of science-driven metadata standards, then the burden imposed on scientists to label their data and submit to data repositories is not warranted.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Response: The DOI (digital object identifier) for each publication should be included in the metadata associated with each data set and conversely, the location of the data should be provided in each publication.
We, and our data sharing task force, look forward to working with OSTP in the future.

Sincerely,

John Fitzgerald
Policy Director