The Ornithological Council

PROVIDING SCIENTIFIC INFORMATION ABOUT BIRDS

American Ornithologists' Union

Association of Field Ornithologists

CIPAMEX (Sociedad para el Estudio y Conservación de las Aves en México)

Cooper Ornithological Society

Neotropical Ornithological Society

Pacific Seabird Group

Raptor Research Foundation

Society for the Conservation and Study of Caribbean Birds

Society of Canadian Ornithologists/ Société de Ornithologistes du Canada

The Waterbird Society

Wilson Ornithological Society

Ellen Paul
Executive Director
5107 Sentinel Drive
Bethesda, MD 20816
Phone (301) 986-8568
Email: ellen.paul@verizon.net

12 January 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 Ornithological Council, a consortium of twelve scientific ornithological societies in the Western Hemisphere, 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.

Ornithology 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 avian biology.

Our organization strongly supports the concept of archiving and sharing these data. We have investigated and discussed the possibility of developing an archive for the types of data generated in ornithological 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 reasonable time after the grant period has ended, so as to allow them to complete their publications. 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 also 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 Center 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 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.

There is already ample evidence that federal funding does result in the development of successful data repositories. Federal funding was largely responsible for the development of a suite of taxonomic databases – ORNIS for birds, MANIS for mammals, HERPNET for herps, and FISHNET for fishes – each a distributed database and all interoperable, mappable, and publicly available.

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.

(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?

Response: Assure that researchers have a reasonable time after the completion of data collection or the end of the grant period, whichever is later, to publish the results before making the data publicly accessible. There may be some situations that merit a longer period of exclusive access. 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 and possibly over several generations. 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. A set of criteria to determine when such extensions are appropriate is needed.

(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?

Response: 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. Further, while it may be that among all the existing repositories, a researcher could find suitable repositories for some parts of the data in a given dataset, it is not reasonable to expect a researcher to have to submit data to two or more different datasets, particularly as it is possible that the two datasets may not use the same metadata standard.

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. Most of the birds banded are 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 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. The same concerns would pertain to location data of other animals or plants protected under the Endangered Species Act, and to other animals that are vulnerable to disturbance, should location data be made available. Even species that are not endangered (whether or not legally protected but that are in commercial demand could be over-exploited and small populations could be driven to extinction by over-collecting. In cases such as these, the researcher should be permitted to omit, obscure, or coarsen the location data.

(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?

Response: It is not clear that this can or should be done. Suppose that the number of queries or data retrievals were reported by each repository? That information would not tell us if or how the data were used, and of course, the determination of the value (benefit) of that use is subjective and not comparable across disciplines.

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

Engage the scientific community full in every stage of the development of data management plans. The NSF data management plan requirements, flexible though they may be at this time, seem to have been developed by the National Science Board without any, or any significant, input from the scientific community. There were two workshops – one in 2003 and another in 2004. At the first, only two hours were allotted to discussion; at the second, only 45 minutes. A relatively few public comments – most from other federal agencies, data management firms and professionals, and only a few from researchers, research institutions, or scientific societies – were received in response to a 2005 request for comments. Between the task force recommendation in the 2005 report and the actual development of the NSF data management plan requirement that went into effect in January 2011, there seems to have been no opportunity for input from the scientific community.

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

Response: 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, but for the support afforded to efforts such as VertNet.

(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?

Training is probably the key to improving compliance. The existing biological data profile and the numerous metadata entry tools that exist are not, in many aspects, intuitive. It is likely that scientists who have not had training will struggle to use these systems and will either give up entirely or will not enter complete information. Training is likely to reduce the barrier to use of the metadata entry tools.

Verification could be a step in grant close-out.

(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 "...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. However, the use of the data in a subsequent analysis is the purpose and benefit of public access; the lack of attribution is not a sufficient reason to curtail access. The real value of the data to the original researcher is that researcher's own publications; the unattributed use of the data in a subsequent analysis does not diminish the value of the original publication or of the use of the data for that original publication.

Standards for Interoperability, Re-Use and Re-Purposing

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

(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 our discipline (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 ornithological 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. The termination of NBII may also result in the termination of funding for the single coordinator position and a single programmer position for VertNet. The participating

institutions are all suffering from the economic downturn and cannot readily replace the funding for these two positions.

(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 any mandate 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 Global Biodiversity Information Facility has developed a protocol for the use of universal identifiers that can be used to refer to a range of digital objects in data sets, documents, and repositories. The single identifier would allow the user to retrieve all digital objects (such as datasets) associated with a publication or, conversely, all publications associated with a dataset. The use of a universal identifier also facilitates the tracking of digital object retrieval and, should that item be used in a subsequent publication, could also help determine the extent to which that information was actually used.

We thank the OSTP for considering our concerns and views, and hope that this response will prove helpful in shaping federal policy on public access to digital data.

Sincerely,

Ellen Paul
Executive Director