Summary Report: Workshop to Address Changing Practices Around the Publication of Biological Data

OVERVIEW

The American Institute of Biological Sciences (AIBS), with support from the National Science Foundation (NSF), convened a one-day workshop on 3 December 2014 that explored the implications of changes in data management practices expected to result from recent and forthcoming federal policy changes. The changes will require public access to most scientific data resulting from unclassified federally funded research.

The event was held at the Capital Hilton in downtown Washington, DC. The workshop consisted of four panels in which 22 panelists representing key stakeholder organizations discussed where data should be published, when there might be valid exceptions to a general mandate, procedures to ensure proper professional credit for data producers, requirements to review data for publication, and sources of funding for the necessary work. Open discussion involved representatives of the journals of a range of biological societies, as well as scientific publishers, federal government officials, and researchers and representatives of non-profit organizations that work in biology.

There was no apparent dissent from the general proposition that more data-sharing would be beneficial to science and society. The workshop focused on identifying productive ways in which journals and other organizations could encourage more sharing of data. This summary report records the comments that were made about obstacles as well as possible solutions; however, no attempt was made to reach a consensus statement and so none is provided in this report. Some suggestions that seemed to command general assent are listed at the end of the report. A draft of this report was published online in March 2015 and comments were invited from all participants; comments received are reflected in this final version. The agenda of the workshop, including a full listing of attendees, moderators, and panelists, is provided in an appendix, but responsibility for the report's content rests with the AIBS staff member who drafted it.

BACKGROUND AND BENEFITS OF SHARING DATA

Since 1945, federally funded research has been the base of the research enterprise, and the understanding has emerged that investigators funded by the government would provide resulting data to the sponsoring agency but would retain ownership rights. Such data are not federal records. The Shelby Amendment of 1999, which was prompted by a controversy over public access to Environmental Protection Agency data on particulate air pollution, stipulated that results of federally funded research should be made public, and resulting administrative changes meant that some data relevant to regulation became subject to the Freedom of Information Act. Most research data do not fall into this category, however.

The memorandum on public access to research publications and data released by the White House Office of Science and Technology Policy in February 2013 sets forth the administration's position on scientific data: digitally formatted scientific data resulting from unclassified research supported wholly or in part from federal funding should be

stored and publicly accessible to search, retrieve, and analyze. A few months later, an executive order made open and machine-readable the default requirement for government-owned data, including scientific data; much of this will be made available via data.gov. These regulatory changes lack the force of statute, however.

Some of the data to be made public will, the administration asserts, drive economic growth. Data publication can increase the statistical power of comparisons and expand the scope of science by allowing, through the re-use of data, informative comparisons that would otherwise be impossible. It can also stimulate the development of tools to make re-use and sharing of data easier. Currently most scientific data are not available for confirmatory analysis, reuse, and repurposing—a situation that the administration seeks to change. The policy will not represent a significant change in those fields of science (such as genomics) in which publication of data is already the norm. But current data management plans in other fields have often included data publication as a "check-the-box exercise." The new norm will be that researchers will be asked explicitly how and where they will share their data.

A long-term administration goal is the development of federated systems of databases that will allow the storage, discoverability, reuse, and repurposing of data and provide data services—a "research data commons." This would allow the discovery of datasets from publications and vice versa. This desired endpoint would have data being a "new currency in science." Researchers would get credit for datasets, not just for publications.

Science has been built on competition for limited funds, the idea being that the best science will rise to the top and be funded and published. In some ways this has been very effective. But it is arguably in conflict with the goals of being open and sharing, which creates a dilemma: Many modern questions and problems can be addressed only through collaboration, but if all data are published, how can funders identify where to allocate scarce resources?

TYPES AND AMOUNT OF DATA TO BE PUBLISHED

There was general recognition of the need to be thoughtful about the level of data that should be shared, and that this could vary immensely between fields; this was at the heart of the controversy over air pollution data that led to the Shelby Amendment. But standards need to be drawn carefully, as they could otherwise stifle research; most comparable standards have to be revised every 3 to 5 years. One participant argued that the desiderata of reproducibility and reuse will often suggest different answers to the level of detail that should be retained in published data.

Different fields of science vary greatly in the type of data that they produce and in their typical practices of data publication. This means that there are major unresolved questions about the technical requirements and the infrastructure needed for long-term data storage in some fields. There are also unresolved questions about the needed procedures and balances for securing adequate public access. Deciding when data can be deaccessioned remains a further difficult problem. The NSF, for one example, is sensitive to the variety of practices and of data, and is striving, while advancing the federal policy

initiative, to learn about best practices from the relevant communities. The NSF adopts a consensus-based approach on publication practices, aiming to foster intellectual creativity. NSF has required a data management plan as part of proposals since 2011. Publication and data preparation charges can be included in budget requests to NSF—although they are deducted from the grant totals—and data publications can be included in investigators' biographical sketches, but existing policy lacks specificity.

NSF may retain most if not all of its current practices, but will look for guidance from program divisions and directorates on extending the public-access directive. Investigators will be able to provide feedback on proposed changes in practices via Web-based systems. [Note: the NSF's Public Access plan was published on 18 March 2015: see www.nsf.gov/news/special_reports/public_access/]

It was suggested that AIBS, as a meta-level organization representing a wide variety of biological societies, might help by developing guidelines on consistent formatting of biological data, because there are divergent understandings of how this should be done. In general, the promulgation of reporting requirements for researchers, such as the life sciences reporting guidelines and checklist used by *Nature*, can increase the value of data.

IMPORTANCE OF FUNDER MANDATES

Almost all agreed that funder mandates are critical to bring about change. One participant told how, when serving on an NSF panel, she had been disappointed that "terrible" data management plans in some submitted grant proposals were of no apparent concern to her fellow panelists.

The National Institutes of Health (NIH) has a history of establishing a data-sharing culture that goes back to the late 1990s. Its policies are effected through research tools, extramural grants policy, and intramural rules on large database sharing. There has been a data-sharing expectation for grantees awarded more than $500,000 in direct costs in any single year per year since 2003. There has also been a policy requiring sharing of model organisms and related resources, including data. More recently, there have been policies on sharing genome-wide association studies and now genomics data. The Office of Human Research Protections in the Department of Health and Human Services has proposed changes to the "Common Rule" (the Federal Policy for the Protection of Human Subjects) to support the maximum utility of specimens and data. The Database on Genotypes and Phenotypes (dbGaP; see Table, pages 3 and 4)), a controlled-access data repository that makes data available under terms and conditions consistent with informed consent provided by individual participants, houses genotypic and associated phenotypic data; investigators get approval for requests to access the data. The NIH Big Data to Knowledge initiative and its associated discovery index will further encourage a data-sharing culture and incorporate tools for measurement. It will thus yield a data ecosystem that supports discovery; this initiative includes the notion of a "research data commons" and a data discovery index.

Table: Selected entities referred to in this report, with brief descriptions and URLs

ENTITY	DESCRIPTION	URL
BD2K (Big Data to Knowledge) initiative	Initiative to overcome impediments to the use big data for understanding health and disease	http://bd2k.nih.gov/about_bd2k.html
Clearinghouse for the Open Research of the United States (CHORUS)	Not-for-profit public-private partnership of scientific societies and publishers working to increase access to federally funded peer-reviewed research	www.chorusaccess.org
Council of Science Editors	Editorial professionals working on effective science communication	www.coucilscienceeditorsn.org/
CrossRef	Association of scholarly publishers that develops infrastructure for cross-linking scholarly communications by citation linking	www.crossref.org/
DataONE	Partnership developing a distributed cyberinfrastructure for earth observation data	www.dataone.org/
DRYAD	Curated general-purpose repository for a wide variety of data types	www.datadryad.org/
Figshare	Commercial data archive and publication portal	www.figshare.com
FORCE11	Scholars, librarians, funders, and others working toward improved knowledge sharing and creation.	http://www.force11.org
FundRef	Public registry of research funding maintained by CrossRef	www.crossref.org/fundref/
GoMRI (Gulf of Mexico Research Initiative)	Program established by BP to study effects of oil spills	www.gomri.org
GRIIDC (GoMRI Research Information and Data Cooperative)	Public database established by GoMRI for all its research data	https://data.gulfresearchinitiative.org/

Table (contd.) : Selected entities referred to, with brief descriptions and URLs

ENTITY	DESCRIPTION	URL
iDigBio (Integrated Digitized Biocollections)	National resource for advancing digitization of biocollections	https://www.idigbio.org
Orcid	Registry of unique identifiers for researchers	http://orcid.org/
SHARE (SHared Access Research Ecosystem)	A higher education and research community initiative to ensure the preservation of and reuse of research outputs.	www.arl.org/focus-areas/shared-access-research-ecosystem-share
VertNet	NSF-funded project to make biodiversity data available on the Web	www.vertnet.org

The traditionally liberal stance of federal agencies on intellectual property (IP) is underscored by the provisions of the Bayh-Dole Act. But the law creates a loophole that makes it difficult to require the public release of data—investigators can claim an exemption from publication requirements if these threaten their IP rights. Publishers have pushed for public release of data in some specific areas, such as with protein and sequence data, linked genotype and phenotype data, and macromolecular and crystallographic data. But there are still unresolved technical issues around the publication of some data (for example, some types of genetic data). There may be a need for more bioinformatics computer programming training for biologists, so that they can themselves better script and convert data between formats.

PROFESSIONAL CONCERNS OF RESEARCHERS OVER DATA-PUBLICATION MANDATES

Reporting on concerns that some researchers have expressed about data publication, one participant noted that requiring public access to data would also take up a lot of researchers' most valuable commodity, time. Researchers spend 40 percent of their research time on administrative duties, so it is important to ensure that only what is important is preserved. When the Public Library of Science (PLOS) announced in early

2013 that data supporting all papers submitted to its many journals must be publicly available, a significant portion of the online scientific community was critical. (PLOS has retained its policy and continues to be one of the world's biggest scientific publishers, however).

One key concern as data-sharing becomes more common, it was reported, is the appropriate attribution of professional credit to data producers as well as to researchers who interpret them scientifically: tracking the origin of data through what may be multiple re-uses becomes important. Some researchers are not convinced they will get enough professional credit from a data set to justify a huge effort publishing it. Data descriptors---short articles about a dataset---may be part of the answer, but researchers may be unfamiliar with having such data descriptors rigorously peer-reviewed. However, a generational shift may be underway: one journal editor stated that his young faculty colleagues are now "enthusiastic" about publishing data, usually linked to from papers, and the number of journals publishing data descriptors is increasing.

Attribution is also a major concern for researchers at iDigBio, a coordinating center for an NSF-funded program for advancing digitization of biodiversity collections. iDigBio was designed to address the problem that information about biodiversity was not flowing adequately to researchers, because very little of it is currently in digitized form. The NSF supports 13 thematic collections networks, and iDigBio coordinates them and ensures their information is made accessible online as efficiently as possible in georeferenced formats. Currently there are 24 million specimen records in the system and there may be a billion in due course. Traditionally, formally publishing definitive information about a species has been a key to academic tenure and promotion in some fields; that professional recognition could be lost if data become public without their originator being identified.

A participant from the Ecological Society of America urged that publishing data be seen as an ethical issue. Many but not all repositories provide a DOI (Digital Object Identifier) for every dataset; in some fields, different identifiers are more suitable. Journals might revise their instructions for contributors to require acknowledgement of data providers as well as mere citation. Participants thought that the use of identifiers, especially resolvable identifiers, is a necessary condition for adequately characterizing and tracking a biological collection, for example, and such identifiers are becoming commonplace in the data world. Automation can improve their usage. Assuring that professional contributions will be recognized can thus facilitate compliance for publication mandates. FORCE11 (www.force11.org) has produced work on relevant standards.

Researchers in fields such as ecology and behavioral science will often exploit the same data set for years, so some have made the argument that requiring public access to data could put such researchers at a particular disadvantage. The argument is also made that poorly funded researchers, including many from developing countries, will be particularly affected by a general data publication mandate. Another objection was that requiring data publication before appropriate standards have been developed will lead to a chaotic profusion of formats that could, in the worst case, make acquiring new data easier than trying to find and understand existing data. In some fields, standards are still lacking; in others, they exist but are not widely used.

Model organism geneticists are also sometimes anxious about being scooped if they publish their data. In addition, they worry that publishing a DOI for a dataset will leach citations away from their articles. Until recent years, researchers wanted to save data for their major publications, so it was sometimes hard to get them to provide it for a lesser paper. This reluctance, albeit now waning, points to the need for systems to guarantee proper professional credit to the original sources of data when data are reused. (Embargoes on data reuse could be another solution).

One suggestion is to involve the original creators of data in reinterpretations and to foster a peer review culture that recognizes the value of new data even when similar data have been published previously; some journals are doing this explicitly when they publish negative results. Telling stories that demonstrate the value of data publication, which is still unclear to researchers in some fields, could be another important way to persuade researchers that the benefits to science and society should outweigh narrow professional concerns in an ethical calculus.

Some users of the Long-Term Ecological Research (LTER) Network do not like that researchers cannot know who is using their data. But because the data are long-term, people have a vested interest in continuing a good relationship with the network, which can track the provenance of all data it delivers. All data packages are subject to quality reviews and receive a DOI, and the system tracks versions. Users and providers of data overlap to a large extent, which is perhaps another reason why the LTER Network has made progress toward data sharing. Still, in general, researchers face an uneven reward system for depositing data, and although most are willing, occasional unethical practices can discourage sharing. This seems to emphasize the need to promote professional best practices.

A participant from the American Society of Plant Biologists was keen to see more education about data sharing generally, because many of the available tools, built by computer scientists, are not very usable by biologists. This, again, suggests a need for a re-envisioning of part of the educational and workforce training system to encourage the inclusion of this kind of training in biology programs.

Others commented that such training was already happening in some areas. One participant recommended “research sprints,” intense 2- or 3-day markup sessions, to encourage data sharing. Another commenter made the suggestion that young scientists should be assigned a specific data repository for their use; this policy could benefit, in particular, marine laboratories and field stations, whose data often “walk away” when visiting scientists leave. It was suggested that educational resources for editors might be developed by the Council of Science Editors. The US Forest Service’s Data Archive (http://www.fs.usda.gov/rds/archive/) has data specialists who help researchers publish useful, well-formatted data, a model that might work elsewhere.

Participants suggested in different ways that the research community should engage in expanded discussions and interactions to promote and encourage the sharing of data,

stressing the value of the practice for science in general and that those who do not share data are not abiding by scientific norms. A related view was that the value of data sharing, while superficially akin to a science value such as the need for informed consent for research on human subjects, has not yet achieved such widespread acceptance; the policy discussion has, hopefully temporarily, gotten ahead of the values discussion.

COST OF DATA PUBLICATION MANDATES

Cost is a major issue in expanding data publication generally. Policing publication mandates will have to be paid for, and production of metadata may be expensive and challenging in fields where data publication has not been the norm. Businesses may not be keen to increase prices to implement an unpopular policy. CHORUS is developing an information system for ensuring compliance with article public access mandates that rides "on top of" the existing systems FundRef, CrossRef, and Orcid. It might at low cost be able to satisfy data compliance mandates too. The Association of American Universities has an initiative known as ShARE (Shared Access to Research Ecosystem). Initially conceived as having a function similar to CHORUS, ShARE has shifted somewhat in emphasis and is now collaborating with a variety of other projects, including CHORUS, with the goal of making all scholarship and supporting data freely available. ShARE has not yet established all its policies on data publication, however.

One participant urged the Federal government to pay for the necessary work with new money, as the government has used taxpayer funds to support research. One participant argued that researchers will not want to use their budgets for data publication, leading to the thought that institutions may have to take on supporting data repositories as they have supported libraries, even if that means cutting library staff. But universities might be targets for exploitation: Data publication initiatives now being funded by foundations will not be sustainable, so universities will have to fund them. And commercial companies might text-mine information published at some cost by universities and exploit it for profit.

The funding question is far from being just hypothetical: databases for model organism genetics exist, but their funding is being cut. Some such databases have tried to change their business model but that hasn't worked well. The Arabidopsis Information Resource almost disappeared two years ago because of financial problems. One participant suggested that researchers should remind NSF and the National Academies of the importance of scientific data centers, where large specialized data sets can be managed. Another suggested that adopting the alternative, Darwinian, approach, will mean that data are lost when repositories fail.

Even large initiatives such as the LTER Network are looking for answers on long-term financial sustainability: The network will transition during the next 18 months to a maintenance status, and new funding is being sought. The Gulf of Mexico Research Initiative (GoMRI) is talking to major corporations, including Google, about a possible private sector solution for long-term storage for its large GRIIDC (Gulf of Mexico Research Initiative Information and Data Cooperative) database.

Fees provide for long-term sustainability of the DRYAD repository. Individuals are charged $80 per paper, and journals that have an exclusive arrangement with DRYAD might pay only $25 per paper. Still, DRYAD has prepared contingency plans for if it should ever have to go out of business.

DataONE—an NSF project that supports cyberinfrastructure development and community engagement for the biological, environmental, and Earth sciences—has concerns about low budgets provided for data management projects, according to its Principal Investigator. DataOne aims to address the loss of environmental data over time; some studies indicate that such data have a half-life of less than 20 years, and although there are many data repositories, there is no complete list. With a backbone consisting of three nodes, DataONE is now replicating metadata catalogues from participating institutions and organizing them into a rapidly searchable database, emphasizing discoverability.

ROLE OF JOURNALS

Publishers of scientific journals have, by common agreement, an important role to play in influencing data publication practices. In genomics, astronomy, and economics, journals require the deposition of data as part of their review process even before publication.

From the publishers' side, ensuring compliance could be a challenge. Reliance on peer reviewers, an informal investigation published in *Nature* suggests, may not effectively ensure compliance with a publication policy.

The objections some researchers raised to the PLOS policy illustrate that there is a need for absolute clarity about data publication requirements. But journals should consider beefing up their requirements on materials and methods sections of articles too, which could speed science appreciably. In many articles they are currently vestigial.

The American Association for the Advancement of Science (AAAS) set data policies for its three [now four] journals in 2011. It lists acceptable data types for authors but identifies suitable repositories only when it believes that these have been accepted by the research community. In other cases, supplementary material or, rarely, authors' own institutions may be accepted as a suitable home for data. Even so, editors still sometimes have questions about what constitutes raw data. AAAS plans to watch which agency Data Management Plans are accepted for funding and to use this information to guide the refining of its own policies.

There are complications, however; some Data Management Plans appear to assume the existence of third party resources (i.e. journals) for hosting data, which introduces an element of circularity, especially if the journals are not aware of this role being thrust upon them. AAAS has convened workshops (funded by the Arnold Foundation) on reproducibility that have looked at data sharing, as well as focus groups, to get community input. AAAS plans to work in conjunction with other journals to set common standards as researchers move from minimal standards to aspirational ones. AAAS is

rethinking its workflows to make upload of data easier and to ensure their accessibility (via DOIs) in published products. AAAS hopes also to make better use of tagging of reviewer comments on data so these could be assessed to calibrate community standards.

One view was that journals have a primary role in ensuring proper citation of data identifiers in publications, not in hosting data, and that new resources should go to curated databases that are publicly accessible, rather than to journals. Editors need to know which databases are stable and curated databases, rather than assume curation responsibilities themselves; databases have more domain-specific expertise than do journal editors. The Protein Data Bank, which curates protein structures, provides a gold standard for data quality. The required use of a curated database can be acceptable to researchers: authors submitting to the *Journal of Biological Chemistry* overwhelmingly agreed with the suggestion that they be required to submit validation reports from the Protein Data Bank. Genetics Society of America (GSA) journals introduced a mandatory data deposition policy in January 2010. Although they lost some papers at first, objections have largely evaporated.

Another argument that journals should not be in the business of storing data is that data are usually not very discoverable as "supplementary material." However, there are different routes to ensuring quality. *Radiation Research* has an in-house statistician to examine analyses in submitted articles to ensure their quality.

And some journal publishers are getting into the business of publishing data, as Macmillan's investment in Figshare, a commercial data repository and publication company, exemplifies. This could be part of the solution if for-profit businesses are held to the same standards of quality as other entities.

Genome Announcements, a new, open-access journal, might be a model solution for some of the difficulties. It publishes short manuscripts announcing the availability of newly sequenced microbial genomes. The American Society for Microbiology (ASM) wanted to improve discoverability and awareness about data on microorganisms. Initially, short articles describing how and why a microbial sequence was obtained were published in the *Journal of Bacteriology*. The articles could not discuss any hypotheses. But researchers became concerned because these little-cited but important articles were dragging down the impact factor of the journal. Accordingly, ASM decided to create a journal just for these announcements. The journal now performs a valuable public service. Articles are not normally sent out for peer review. Young researchers are very open to data-mining explorations and see value in such publications, even though they do not fit well into the more traditional paradigm of hypothesis-driven science.

The American Society of Naturalists (ASN) has a data policy mandate for its journals that dates to February 2010 and is "loose by design." Some communities were willing to provide data, but ecologists, evolutionary scientists, and behavioral scientists were very reluctant. Some argued about the granularity of the data that might meet the policy mandate, objecting to having to include individual replicate values. Others objected to providing the raw numbers coming off machines, but were willing to supply them after

the information was processed and massaged. The ASN decided to let authors decide how to meet the data mandate.

ASN is uncertain whether to require the deposition of metadata pertaining to duplicate data sets; laboratories have somewhat different procedures and some differences between datasets might therefore be artifacts. Particularly problematic are duplicate datasets that are deposited at different levels of "massaging": Most researchers want to examine only the version of a dataset that they were specifically referred to, and baulk at tracing it back to its original source. ASN has a small staff that cannot examine deposited data in detail and so must trust researchers.

Evolution, published by the Society for the Study of Evolution, does require data deposition: the policy is that it should be possible for another investigator to recreate every result. The journal has seen no falloff in submissions since it put the policy in place. But it does not check the quality of deposited data, and the necessary annotation (provision of metadata) can involve authors in a lot of work.

The LTER network has successfully arranged for its repository to be accepted by a number of journals as a suitable archive for published data, but there is still no consistent approach to citation of data or clear message that it should be done. It was suggested that this would be a good goal for the workshop to try to achieve; there needs to be long-term planning.

It may be desirable for journals to encourage researchers to share their data and show them how to do so, since many are not familiar with the needed software and systems. *Gigascience* has successfully published impressively large datasets and provides a test platform where authors can rerun analyses on large datasets.

Journals might even suggest that data providers are made, in some circumstances, co-authors of an article that re-uses their data. Sharing might be encouraged in codes of ethics. However, there are reasons for caution about providing authorship to data providers, because authorship implies responsibility for the content of an article; data providers could find themselves responsible for ethical problems in an article that reused their data.

POSSIBLE MISUSE OF DATA

Some participants expressed concern about actual and possible misuse of data. In some cases researchers have cheated on a data publication agreement by modifying data to make them less useful to others after publication of their paper. One editor knew of papers that are based on fraud that are still being cited. It is the responsibility of data users, therefore, to ensure the validity of data that they re-use. It is not far-fetched to imagine that a bad actor could deliberately upload misleading data to a repository. On the other hand, the NIH is aware of cases of researchers claiming credit for papers for which they provided data as their own even when they have made no intellectual contribution.

SHOULD DATA BE PUBLISHED OR MADE AVAILABLE TO REVIEWERS BEFORE PUBLICATION OF AN ASSOCIATED ARTICLE?

One important question is: When exactly during research practice and reporting do data cease to become the personal property of a researcher and come under the control of a publisher or the government? Making data available to reviewers before publication of an associated article clearly has the potential to improve the usefulness and accuracy of data publications, although there is evidence that in some fields, at least, reviewers do not pay as much attention to data made available as might be hoped. Embargoes on data re-use can allay researchers' fears about being scooped; several journals allow such embargoes.

Attempts to encourage data publication within the University of California system ran into trouble at the office of the University president's legal counsel, who wanted to make a claim of ownership on data produced within the system and so baulked at releasing them with a CC 0 license. The university has accepted mandates to allow public access to scientific articles, but data present many more difficult issues, because copyright does not apply to them. Moreover, data are complicated by privacy issues and the potential for data to be valuable if sold, for example to pharmaceutical companies. Publication of computer code is another very complex matter on which the university is still trying to arrive at a satisfactory policy.

Scientific societies and organizations may be able to help answer questions of ownership, given the large differences in practice between scholarly disciplines, because they provide a structure that can facilitate decisionmaking. Currently, policies vary; allowing anonymous review of data in repositories has the potential to allow prepublication review of data to become more common.

The LTER Network generally requires the prompt deposition of all data generated in its system of research sites; the only exception is that graduate students are allowed to set an embargo on data publication for longer than others.

Gigascience encourages release of data before publication of an article, although it allows researchers to put an embargo on the time of public release of the data. *Gigascience* provides a database for data types that lack a community repository. *Gigascience* also proactively helps researchers gain access to protected data (such as that relating to human subjects in dbGaP) by making the required release forms available to them.

The Botanical Society of America's (BSA's) journals have long required that data associated with articles to be published before final acceptance of manuscripts. They see tagging of data in a repository as a better solution than publication of supplementary materials. Supporting vouchers and accession numbers are required with all submissions. Excuses based on anxieties about being scooped or the time needed to provide the data are not generally accepted, although exceptions are made on rare occasions, for example to avoid publicizing the locations of endangered species. In the future there may be much more reciprocal linking via citable DOIs, and BSA journals will shortly institutionalize their data practices by requiring a data accessibility statement from submitting authors.

The editor in chief of *Functional Ecology*, who is a member of the publications committee of the British Ecological Society, said that earlier opposition to data publication in ecology had disappeared over the past five years. But many researchers still opposed making data available for review before publication of an associated article, especially when they thought that their article had only a small chance of being accepted.

AAAS sees DRYAD as a possible solution for reviewers' fears that their identities might be discoverable by authors if data under review are hosted on an institutional website. Reviewers can examine data hosted in DRYAD without compromising their anonymity. Yet one editor pointed out that even if data review is made anonymous by depositing the data in a repository such as DRYAD, researchers still worry that reviewers who may be competitors will see their data.

Editors of some journals have been reluctant to make data deposition a condition of submission, in part because they may not yet "have a handle" on all the types of data that will be submitted. One compromise, adopted by *Elementa*, an open access journal published by BioOne, is to require authors to answer the question "What have you done with your data?" *Elementa* does facilitate deposition of data to DRYAD. Editors can check that data have been deposited before a publication "goes live" online, although *Elementa* does not require peer review of the underlying data and authors may request an embargo on data publication.

At the GSA, editors do make basic checks about claims that data are deposited but do not always check exhaustively; in one instance a controversy resulted when authors of a paper withdrew referenced sequence data on the HeLa cell line after the paper was published. Some controlled access to the sequence data was restored for researchers after an intervention by the director of the NIH.

Journals of the GSA do allow the right of first use to data providers, that is, an embargo on the publication of their datasets. ASN has accepted requests to embargo data used in long-term studies but has not yet established a policy for deciding how long the embargo should be (one group of researchers asked for a 60-year embargo on data for a long-term demographic study). ASN does, however, require that data, once deposited, must remain so permanently and cannot be removed. It is still wrestling with what deposition should be required for computer code.

EXAMPES OF SUCCESSFUL DATA-SHARING INITIATIVES

Some special-purpose research organizations have successfully created data-sharing cultures. Accounts of their practices illustrate some of the obstacles that may arise with increased data-sharing, as well as solutions that have been effective.

GoMRI, which was funded by BP about 30 days into the Macondo spill in 2010, conducts research on oil spills, with an emphasis on the Gulf of Mexico. The data are all uploaded into GRIIDC at the Harte Institute at Texas A&M University. That group advises funded researchers on how to accomplish the GRIIDC data policies, which mandate publication of data by the time of publication of an article or within a year of

being collected. Much attention is paid to metadata formats and DIFs (dataset integration forms); consistency of vocabulary is important. GRIIDC provides a summary to the GoMRI board every four months.

So far the effort has been very successful, and research groups compete to complete their data uploads: nearly a petabyte of data has been deposited. The system is improved by suggestions from users, and data portals are linked to scientific publications. Data are georeferenced and searchable by keywords to facilitate discovery. Because data publication is a new expectation for some researchers, some are slow to "flip the switch" to make data public. But compliance is good because uploading data is a criterion for the provision of new research awards. Server size is a continuing issue, however: GRIIDC has not yet migrated to the cloud.

VertNet provides accessible biodiversity data for vertebrate species. The system is based on Darwin Core standards and is a cloud-based aggregation. It takes the place of the formerly separate FishNET 2, MaNis, HerpNet, and Ornis systems. It has a cache so results are delivered quickly. Many technical challenges had to be overcome in the creation of VertNet, as it is integrating data from multiple sources: 182 institutions have contributed data from 509 collections. It is a powerful tool for research in specific geographical areas, for example. It is now being extended to observational data that were not available online until recently, for example observational data on whales and turtles.

The LTER Network uses a flexible metadata approach to allow open sharing of its complex and varied data. This was judged superior to a standardized data structure approach. Each node in the network has its own data manager and the network follows and measures use and reuse of all its data. Communities of practice define the best practices used, and the network is reviewed every three years by a panel that includes an information manager, who checks that data are available throughout the system.

One big issue for iDigBio is handling corrections. In taxonomy, names are quite often changed, and many specimens are misidentified. Often, data-supplying institutions say they don't have the resources to make needed corrections, which is becoming a problem. This emphasizes the need for proper versioning. A wider role in research and education will, however, be the result of continued progress.

DataONE is addressing technical obstacles to data reuse by making large amounts of data easily accessible to researchers. It can direct queries to DRYAD or to the LTER Network or to the US Geological Survey. It has created a series of one-page guides for researchers on best practices on handling data and metadata more effectively. The work will allow scientists at small institutions to extract data from extremely large data sets. DataONE is also developing tools and workflow packages to allow easy semantic annotation of data. One tool is designed to filter data sets while ensuring the user always knows where the data came from.

DRYAD is a database for varied biological data that is now run by a stand-alone member organization. Its data sets are curated and are linked reciprocally to journal publications.

DRYAD emerged as a solution for experimental data publication. Every dataset gets a DOI, and uploaded data get a creative commons license. DRYAD is often used for peer-reviewers to anonymously review data, but it also hosts many datasets long-term. DRYAD tracks download data and hopes to develop readily accessible information on re-use of its data.

DATA RETENTION

There was general agreement that it is unrealistic and unnecessary for all data to be retained indefinitely. For example, improvements in experimental techniques might make older data irrelevant. NSF allows researchers to state how long their data should be retained in their data management plans. Still, projects such as NEON (National Ecological Observatory Network) will produce enough data to challenge existing policies. Decisions will have to be made on a case-by-case basis: for example, experimental data from a species that has gone extinct should be retained even if the techniques had been superseded. But for some replicated data sets (for example, data created to test instruments) there may be no justification for archiving and publication.

KEY SUGGESTIONS

Although no formal consensus statement was sought at the workshop, the following suggestions appeared to command general support and provoked no apparent dissent. They are listed in arbitrary order.

1. There should be wider discussion among researchers of the expected benefits to science and society from increased data publication, as this will help to clarify the values argument and solidify professional norms favoring publication of data.
2. Journals can play an important role in pressing researchers to publish their data, ideally in a suitable repository, in a timely fashion, allowing only rare exceptions; temporary embargoes on data publication may be justifiable in some cases. Journals should weigh the possible benefits of having reviewers examine data before its publication, as this can be done without compromising the reviewers' anonymity.
3. More work is to needed to standardize data and metadata formats in fields that do not currently have well-developed, widely accepted standards, and to promote the use of existing standards. Standards should be revised every few years to reflect changes in technology and practice.
4. There is a need to develop more user-friendly information technology solutions to allow easy tracking of the provenance of data, including resolvable identifiers.
5. Governments and other institutions that fund research should recognize the importance of ensuring the availability and long-term financial sustainability of data repositories with adequate storage capacity and resident expertise in data formats.

6. Research funders and other organizations should attach a high priority to ensuring the availability of education and training for researchers in the skills necessary for data management and archiving.
7. Institutions should strive for clarity in how their Intellectual Property (IP) policies treat data, so that their federally funded researchers and journals will be able to determine when IP concerns should delay publication of data.

###

Appendix

AGENDA FOR WORKSHOP ON CHANGING PRACTICES IN DATA PUBLICATION, 3 DECEMBER 2014.

8:30 – 8:40 **Introduction**
Richard O'Grady, AIBS. Why AIBS is hosting this workshop.

8: 40 – 10:00 **Panel 1: Federal and National Overview**

Moderator: Paul F. Uhlir, US National Academies
Michael Stebbins, Office of Sience and Technology Policy. Rationale for open data and progress report
Eric Fischer, Congressional Research Service
Amy Friedlander, National Science Foundation. What the agency is doing in response to government-wide policy directives esp. wrt. biology. Need for shared responsibility, community input.
Dina Paltoo, National Institutes of Health. Experience of data publication in the biomedical sciences at the National Institutes of Health and its special concerns

– Data access and science: how can increasing access increase scientific discovery and innovation? (presentations to alert participants to opportunities for accelerating science and improving the integrity of published science)

10:20 – 12:15 **Panel 2: Publishers/Librarians**
Moderator: Steward Pickett, Cary Institute of Ecosystem Studies
David Crotty, OUP and CHORUS
Monica Bradford, AAAS, *Science*
Susan Skomal, BioOne
John Vaughn, AAU
Laurie Goodman, GigaScience
Carly Strasser, California Digital Library

-- State of affairs: what data are now being published? When should journals host data themselves, and when should they require deposition in a repository? The new PLoS policy, new "data only" journals. Concerns (extra workload and storage capability;

Editors' knowledge of technical feasibility and expertise on interoperability; necessity for identifiers and other requirements discouraging contributors) and opportunities (greater use of published content and greater scientific integrity, more incentive to publish negative results).

-- What do publishers need from government, scientific community? How are exceptions to data publication mandates to be confirmed? How is the adequacy of metadata to be defined?

-- Are there financial concerns or opportunities? Are there legal concerns in terms of liability and copyright?

1:00 – 2:45 **Panel 3: Scientific Societies and Journals**
Moderator: Charles B. Fenster, University of Maryland
Tracey Depellegrin, Executive Editor, *Genetics*
Amy McPherson, director of publications, Botanical Society of America
Philip Matsumura , Editor, *Genome Announcements*
Mark McPeek , Rep. to Dryad, *American Naturalist*
Jeff Conner, Handling Editor, *Evolution*
Cliff Duke, Ecological Society of America

-- What do scientists need to move toward more data publication? How are data to be discovered?
--How is professional credit to be conferred and recognized? How do researchers ensure they have expertise to ensure interoperability of their data formats? What do researchers need in terms of technical training and help?
-- What are the concerns (e.g. ensuring data are not misused; could there be legal liabilities; how is work involved in providing metadata to be done)?

3:00 – 4: 20 **Panel 4: Data Services / Providers**
Moderator: Judith E. Skog, George Mason University (Emerita)

Chuck Wilson, GoMRI
Hank Bart, FishNET, VertNET
Robert Waide LTER Network
Larry Page, IDigBio
Bill Michener, DataONE
Chuck Fox, DRYAD
-- How are complex data challenges being addressed (or not)?
-- How is usage reported?
-- Are people participating?
-- How is discovery facilitated?
-- What are the obstacles that could prevent wider publication of data?
--Who is going to pay for the expanded infrastructure that will be needed?
-- What steps could overcome these obstacles?

4:20 – 5:00 **General Discussion**
What comes next?
Are there consensus concerns that have arisen?
Is there a need for another discussion, a model plan? Are there volunteers to contribute to a report for NSF?
What should a working group attempt?

Other attendees (including online attendees): Tim Beardsley, Richard O'Grady, Christine Lowe, Lisette Waits, Larry Schmidt, Christopher Randle, Jennifer Solomon, Allen Dearry, Stacey Standridge, Leigh Finnegan, Michael Carron, David Leslie, Lauren Showalter, Sarah Oktay, Matt Kornfield, Cynthia Parr, Sheri Potter, Stephen Gallo, Robert Gropp, Martha Fedor, Stephanie Morris, George Middendorf, Jerzy Jaromczyk, Crispin Taylor, Steve Hageman, Brian Wee, Neil Thakur, Melissa Cragin, Dave Rugg, Pamela Diggle, Theresa Culley, Beth Parada, Ginny Pannabecker, Mary Molinaro, Brooks Hanson, Andrew Sallans, Roland Roberts, Joan Fuller, Meredith Morovati, James Verdier, Adrienne Sponberg, Raymond Mejia, Richard Mitchell, Marc Mendonca, Marilyn Norconk, Steven Holland, Thomas Turner, James Watkins, Susan Musante, Julie Palakovich-Carr, Eric Nagy, Michael Cato, Alan Savitsky, Teri Balser, Deborah Plavin.

Report prepared by Tim Beardsley, Editorial Director, AIBS Publications.