Mon 1/2/2012 8:07 PM
response to RFI on data sharing

Recommendations on ensuring long-term stewardship and encouraging broad public access to unclassified digital data that result from federally funded scientific research

George B. Moody
Harvard-MIT Division of Health Sciences and Technology Cambridge, MA

I am one of the founding members of PhysioNet (http://physionet.org), an NIH-funded resource that curates and provides free web access to many large collections of recorded physiologic signals and time series, and to related open-source software [1,2]. PhysioNet, established in 1999, is intended to stimulate current research and new investigations in the study of complex biomedical and physiologic signals. About 45,000 visitors use PhysioNet each month, accessing data and software contributed by federally-funded researchers and others worldwide. As a measure of effectiveness, a Google Scholar search returns (as of January 2, 2012) over 7000 articles and patents citing PhysioNet or making use of data or software provided by PhysioNet [3].

The observations and recommendations below are mine alone, although they are informed by personal experience with PhysioNet and with related data-sharing efforts beginning in about 1978. They may not reflect the opinions of my employer.

Preservation, Discoverability, and Access

(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?

When applying for funding, researchers understand that their opportunities to work on the problems they may have invested years of their lives to pursue depend critically not only on the quality and relevance of the ideas they propose to study, but also on their scientific productivity and the impact of their work, performance metrics based on their publications.

Many researchers are understandably reluctant to share data they may have invested considerable effort in collecting, when they have little or no reason to expect recognition (i.e., improved likelihood of obtaining research funding) for having done so. This reluctance may be compounded by apprehension that a competitor for funding may learn something from their shared data that will improve his or her own chances.

It may not be possible to provide a set of motivations for data sharing that can overcome these perceived incentives for data hoarding in all cases.
Funding agencies can take a major step in this direction, however, by directing peer review panels charged with ranking research proposals to consider the applicants' data-sharing record. Agencies wishing to encourage data sharing might require that assessments of applicants' scientific productivity

and impact of their work should reflect the amount, quality, timeliness, and relevance to current research of the data they have shared.

(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?

This question assumes that researchers may have IP interests that need to be protected from damage caused by disclosure of data collected using public funds.
It is not obvious that this assumption is warranted.

(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?

In my field, and in others in which data may include identifiers that permit them to be associated with individual human subjects, privacy concerns are often perceived to trump the public interest in accessing research data. At a minimum, researchers must anonymize (deidentify) data before making them publicly accessible. This process can become expensive, and it is often difficult to determine if all identifiers have been removed.

As a result, data sets that might help to address major public health challenges are often not shared, and agencies fund redundant data-collection efforts that fall short of what might be possible if larger numbers of subjects were studied (by combination of data sets) or if the population entered into a single study were examined in greater depth by other investigators.

Agencies might help by establishing repositories for controlled access to data that have been scrubbed to a high standard that nevertheless may fall short of complete deidentification, indemnifying researchers who contribute data to such repositories, and requiring those who access the contributed data to refrain from any use other than scientific research.

(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?

[no recommendation]

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

Institutions such as university libraries may be particularly appropriate curators of shared research data. They are vital to the educational missions of the institutions to which they belong and from which they receive stable long-term support. Academic researchers can collaborate with libraries to develop new and more effective tools for dissemination of their research data to students and educators as well as researchers. Universities may perceive opportunities to strengthen their research activities by promoting data sharing to build collaborations among their faculty members.

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

First, it is appropriate that funding applicants should propose a plan to preserve the data they will collect and to make them accessible, and that this plan must include a budget for doing so.

Second, funding agencies might wish to consider awarding small supplementary grants for data-sharing to grantees whose data are nominated by others as likely to be valuable if shared, on the basis of the progress reports. This might go a long way towards addressing the objection that the costs of data sharing cannot be carved out of a research budget painlessly. For a typical NIH R01, perhaps an award of $15-25K would be sufficient. Such awards would be especially useful in the case of existing grants for which data-sharing is mandated or desired but not explicitly budgeted for.

Third, it is appropriate that applicants' past performance with respect to data sharing be given significant weight during evaluation of new proposals, just as their publications are considered as evidence of the impact of their work and of their productivity. If the expectation is that research data derived using public funds belong to the public, then those who have shared their data have met expectations. Those whose data have been used by other researchers, as evidenced by secondary publications and letters of support, may deserve an extra measure of credit. Those who have not shared their data, especially after receiving a grant under terms that require data sharing, should not receive additional funding until they have met the terms of any previous grants.

(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?

Recipients of funding are already required to produce periodic progress reports detailing the results of their funded research. It should require very little if any additional effort for grantees to document progress on their data sharing plans.

In my field, it is unfortunately common for researchers to put off efforts at data sharing until the conclusion of the major study, since it is also common to withhold data from other researchers until that time. The all-too-frequent result is that funding runs out, the graduate student who collected the data (and who is the only one who knows how they are organized) has moved on, and the principal investigator is writing the next grant application. Data sharing becomes an afterthought, and if it happens at all it will be late, incomplete, unnecessarily expensive, and sub-optimally organized for re-use. These experiences may not be typical of projects in other fields, but I suspect it is the norm for many of them in which there is not already a well-established culture that expects data sharing.

Agencies can help to steer their grantees away from this counterproductive pattern by encouraging or requiring that data be deposited at the times that progress reports are due, in an archive accessible to the agency. Although I would not expect a funding agency to review all such deposits, they should be considered as appendices to the progress reports, and grantees should expect that they will be examined if questions arise with respect to research progress. Researchers should also be encouraged to share these data appendices with external advisors who may be able provide early feedback on (re)usability.

(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?

PhysioNet sponsors annual open challenges aimed at accelerating research on significant problems that can be addressed using freely available data [4].
These events offer an opportunity for anyone interested to work on a worthwhile question and (perhaps) to make progress towards its solution without requiring the lengthy, difficult, and expensive data-collection effort that would otherwise be needed in order to begin serious work.

(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?

It is already in a researcher's best interest to cite sources of data, just as it is expected that researchers will cite publications that are relevant to new work. Doing so establishes a basis for understanding the innovative aspects of the new work; at the very least it allows one to demonstrate awareness of what has been done by others. Citing a well-known and generally available data source can often answer the otherwise difficult questions that peer reviewers may raise with respect to methods of data collection, selection of subjects, and experimental conditions. In scholarly writing, there is no incentive to misattribute or plagiarize data, because proper attribution only adds credibility to the work.

Nevertheless, in some circumstances, explicit disincentives may be needed to deter data plagiarism or misattribution.

Funding agencies wishing to provide such disincentives might begin by adopting a policy that applicants for funding must cite data sources in all of their publications (or presentations, or grant applications) making use of them, and that failure to do so constitutes scientific misconduct that will impact the offender's eligibility for funding to the same extent as any other fraud.

Researchers wishing to avoid having their shared data be plagiarized can make them freely available, and specifically accessible to search engines such as Google. It then becomes trivial to check for suspected plagiarism, and the high likelihood of exposure acts as a further disincentive to misbehavior, if indeed one is needed.

Finally, to the extent that data sharing becomes the norm in a field of research, papers that report on data that have been neither shared by the authors nor attributed to another source should not pass peer review, and will be judged unreliable and likely fraudulent. Journals may be able to accelerate this evolution of the norm by adopting standards that require (or encourage) data sharing.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and
repurposing of digital scientific data? For example, MIAME (minimum
information about a microarray experiment; see Brazma et al., 2001, Nature

Genetics 29, 371) is an example of a community-driven data standards effort.

MIAME and related efforts (see MIBBI [5]) address the issues of how to describe a data set (e.g., in English) with sufficient detail to permit reuse. In general, these "minimum information" projects are aimed at specifying the content of the description, rather than its format or the format of the data themselves.

There is much less to say about data formats, but there are nevertheless important points about formats to be considered in the context of data standards:

Most important is the use of open formats. Readability of data must not be dependent on availability of a specific computing platform (operating system and CPU) and/or a specific reader application. Ideally, creating a reader for a new computing platform should be a simple process.

When research data must be deidentified before public access can be allowed, it is especially important to avoid the use of proprietary formats that may conceal data identifiers thought to have been deleted.

In my field, data files often contain lengthy time series of observations. Frequently, short intervals ("events") within a long series are of particular interest, hence there are advantages to formats that permit efficient random access. do not require that files be read from its beginning in order to locate a desired time interval.

Finally, one of the lessons drawn from PhysioNet is that the use of common data formats is important. We encourage all contributors to use open formats for their data, in part to ease readability, but also so that researchers who may use contributed data will not need to learn to use a new set of tools for each new data set. This is an important component of the value of shared 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 my field, EDF (the European Data Format for polysomnograms and EEGs [6]) is an excellent example of a highly successful standard for a data storage format. It was designed by a handful of medical engineers who met at a conference in 1987 and published the specification in 1992. EDF is currently used by at least 90 companies, including all of the major manufacturers of polysomnographs and EEG recorders worldwide. The (free) specification fits on a single well-written page, and an EDF reader or writer can be implemented from scratch in a day or less by a competent programmer. The format is reasonably storage-efficient (important since EDF recordings can be quite lengthy).

Another example is SCP-ECG (Standard Communications Protocol for computer assisted Electrocardiography), which was designed between 1989 and 1991 and then redesigned in a formal standard development process between 1995 and 2001 in the US, continuing until 2005 in Europe. The SCP-ECG standard is described in a document of about 200 pages [7]. Embedded metadata selects combinations of many alternative formats included at the behest of representatives of manufacturers who participated in the standard development. Most of the variant formats are highly storage-efficient.

SCP-ECG has not been widely adopted, even by the manufacturers represented on the committee responsible for the decade-long standards process.

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

It is vitally important that digital data standards be open. Standards that are encumbered by proprietary technology divide the worldwide community into those who have the right to use them, and those who don't. The only reason to have a data standard at all is to promote communication. Hence the most valuable advocacy role for Federal agencies in data standards development is to resist the intrusion of proprietary technology into data standards.

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

It is not clear that any action is needed in this regard, but it may help to accelerate what is already a growing trend of citing data sources by mandating such citations when shared data have been used.

References

[1] A.L. Goldberger, L.A.N Amaral, L. Glass, J.M. Hausdorff, P.Ch. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley. "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," Circulation 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/cgi/content/full/101/23/e215]; 2000 (June 13).

[2] G.B. Moody, R.G. Mark, A.L. Goldberger. "PhysioNet: Physiologic Signals, Time Series, and Related Open Source Software for Basic, Clinical, and Applied Research," Proc. 33rd IEEE EMBS 8327-8330 (2011). [PDF attached]

[3] http://physionet.org/publications/

[4] http://physionet.org/challenge/

[5] http://mibbi.org/index.php/MIBBI_portal

[6] http://www.edfplus.info/

[7] ANSI/AAMI EC71:2001