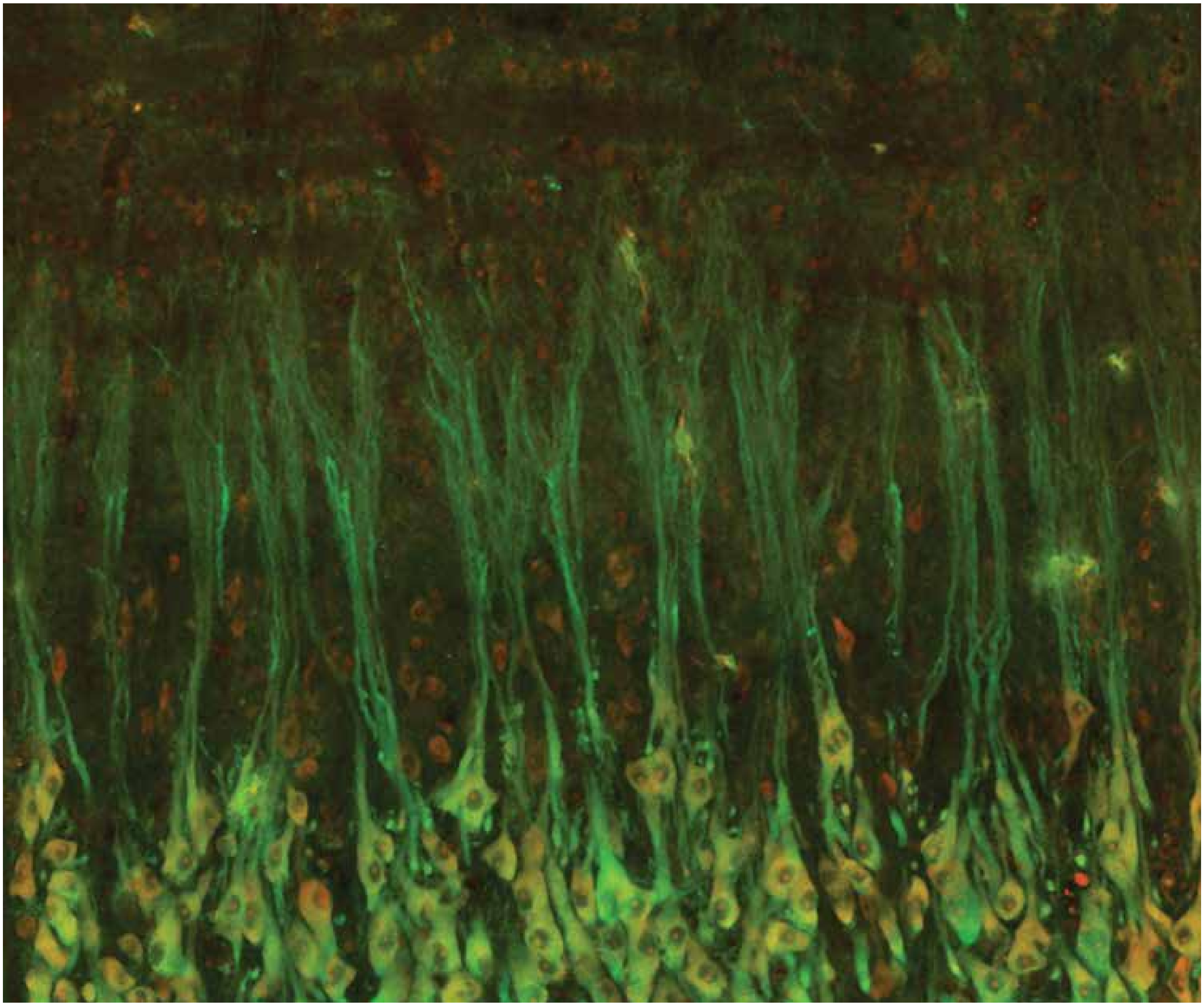

Neuroscience 2016

SHORT COURSE 3

Record Keeping and Data Management for High-Quality Science

Organizers: Michele Basso, PhD; Katja Brose, PhD; Horacio de la Iglesia, PhD; Sabine Kastner, MD, PhD; and Rae Nishi, PhD

SfN Society for Neuroscience

Short Course 3

Record Keeping and Data Management for High-Quality Science

Organized by: Michele Basso, PhD; Katja Brose, PhD;
Horacio de la Iglesia, PhD; Sabine Kastner, MD, PhD; Rae Nishi, PhD

SfN
Society *for*
Neuroscience

Please cite articles using the model:
[AUTHOR'S LAST NAME, AUTHOR'S FIRST & MIDDLE INITIALS] (2016)
[CHAPTER TITLE] In: Record Keeping and Data Management for High-Quality Science.
(Basso, M; Brose, K; de la Iglesia, H; Kastner, S; Nishi, R) pp. [xx-xx].
San Diego, CA: Society for Neuroscience.

Neuroscience
2016

Short Course 3

Record Keeping and Data Management for High-Quality Science

Organized by: Michele Basso, PhD; Katja Brose, PhD; Horacio de la Iglesia, PhD;
Sabine Kastner, MD, PhD; Rae Nishi, PhD

Friday, November 11, 2016
1–5:30 p.m.
Location: San Diego, Convention Center
Room: 11B

TIME	TITLE	SPEAKER
12:30–1 p.m.	**Check-In**	
1–1:10 p.m.	Opening Remarks	Horacio de la Iglesia, PhD *University of Washington*
1:10–2:20 p.m.	Data Management: What Is It and Why Should We Care?	Michael Kalichman, PhD *University of California, San Diego*
2:20–2:35 p.m.	**Break**	
2:35–3:45 p.m.	From a Sponsor’s View: What Data Needs to Be Maintained, Where, How Long, in What Form, and Why?	Ann Hammersla, JD *National Institutes of Health*
3:45–4 p.m.	**Break**	
4–5:10 p.m.	Promoting Open Science: Data Sharing	Maryann Martone, PhD *University of California, San Diego*
5:10–5:30 p.m.	**Question and Answer**	

Table of Contents

Data Management: What Is It and Why Should We Care?
Michael Kalichman, PhD 5

From a Sponsor's View: What Data Needs to Be Maintained, Where, How Long, in What Form, and Why?
Ann Hammersla, JD 7

Promoting Open Science: Data Sharing
Maryann E. Martone, PhD 14

Data Management: What Is It and Why Should We Care?

Michael Kalichman, PhD

Director, Research Ethics Program
University of California, San Diego

NOTES

As researchers, we spend a great deal of time thinking about data as a means to an end (e.g., new insights, publication, or funding), but too often fail to ask fundamental questions about how we should acquire, save, and share our data. The purpose of this presentation is to explore some of these essential dimensions of data management.

The presentation will begin with a focus on six fundamental questions:

1. ***What do we mean by "data?"*** It's clear that recording the temperature of an experimental subject in a lab notebook is a form of data, but it isn't as clear when or if we might consider video recordings, transgenic animals, or software to also be data.

2. ***How do we design and prepare our studies for data collection?*** Experiments need to be designed to not only answer key questions, but, for example, to limit the risk of bias or inadequate preparation of those collecting the data.

3. ***How do we record data?*** If the purpose of research records is to be able to reconstruct what was done for a number of different purposes, then researchers need to consider not only what needs to be recorded, but how to do so in a way that will be useful even years later.

4. ***Have we prepared strategies to address problems that are likely to occur?*** While not all problems can be anticipated, some can. For those likely problems, researchers can and should prepare guidelines and options for what to do when something goes wrong.

5. ***How do we store our data?*** Even the best of research records will have little value if they are lost or stored in such a way that they are not retrievable. Developing and implementing plans for secure storage can at least mitigate that risk.

6. ***How, when, and with whom are we prepared to share our data?*** Although results of research are routinely shared as part of publications, there are many points in the research process in which at least some portion of data or analyses might be shared before publication, and certainly after publication. Responsible data management depends on addressing such sharing.

For each of the previous questions, the goal will be to draw on the collective experience of those in the workshop as well as to identify practices that might be characterized as best. This will be followed, depending on the remaining time, with discussion of one or two representative cases.

Discussion Materials

1. ***Research Records*** Huizhong is a postdoctoral researcher who has worked in the research group of Professor Owusu for the past two years. She has been very successful in her work, publishing several high impact papers. Based on her plans to build on this work, Huizhong has been recruited into an excellent academic position. However, as she prepares to leave, Professor Owusu tells her that she cannot take copies of her research records with her.

   - Can Professor Owusu do this?
   - What options does Huizhong have now?
   - How could this situation have been prevented?

2. ***Ownership*** Nicole and Yuna collaborated successfully for many years, but a personal disagreement has left them unwilling and unable to continue their collaboration. Nicole claims that, because she received the majority of the funding, the data jointly collected should effectively go to her, allowing her to publish without Yuna. Yuna claims that the data should go to her because although she brought in fewer dollars, she wrote more successful proposals for the research and had done more data collection than Nicole.

   - Is Nicole or Yuna correct in her presumption that the data now belong to her?
   - If you were asked to help resolve this dispute, what would you suggest?
   - Is there anything that these collaborators could have done in advance to decrease the risk of this dispute?

From a Sponsor's View: What Data Needs to Be Maintained, Where, How Long, in What Form, and Why?

Ann Hammersla, JD

Director, Division of Extramural Inventions and Research Resources
National Institutes of Health (NIH)

NOTES

Increasing access to digital research data presents significant scientific opportunities to enhance and support its reuse, reproducibility, expand accountability, enhance the return on investment, and accelerate discovery and progress. To seize these opportunities digital data must: (1) be managed and shared appropriately through infrastructures, such as data repositories; (2) be citable to make clear its original source and allow authors of the data to accrue recognition; and, (3) prioritize data sharing activities. In addition, data often must be considered in conjunction with other related digital objects including experimental and analytical workflows, standards, data annotations, and software that act on data. To be effectively shared, data should also conform to the FAIR Principles: findable, accessible, interoperable, and reusable.

NIH's mission is to seek fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability. In addition to funding biomedical and behavioral research into the causes, diagnosis, prevention, and cure of human diseases and the training of basic and clinical researchers capable of carrying out such research, NIH is also responsible for expanding the knowledge base in basic, medical, and associated sciences and ensuring a continued high return on the public investment in research. Sharing biomedical research data is a critical component of the scientific process because it: allows for verification, reproducibility, and validation of findings which enhances quality control of research data; strengthens the statistical power of studies; reduces duplication to improve the return on investment in research; increases transparency in government-funded research activities; and facilitates a scientific collaboration.

NIH has a long history and continued commitment to ensure that, to the fullest extent possible, the results of federally-funded scientific research are made available to the general public, industry, and the scientific community. NIH has maintained the principle that "data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health," and, to that end, has developed a number of policies to further promote sharing of data, such as the 2003 NIH Data Sharing Policy, the 2014 NIH Genomic Data Sharing Policy, the 2015 NIH Intramural Human Data Sharing Policy, and the 2016 NIH Policy on Dissemination of NIH-Funded Clinical Trial Information.

On Feb. 22, 2013, the White House Office of Science and Technology Policy (OSTP) released its memorandum entitled *Increasing Access to the Results of Federally Funded Scientific Research*. The memorandum directs federal agencies and offices to develop and submit plans to OSTP that ensure that peer-reviewed publications and digital scientific data resulting from federally-funded scientific research are accessible to the public, the scientific community, and industry to the extent feasible and consistent with applicable law and policy; agency mission; resource constraints; U.S. national, homeland, and economic security; and, the specific objectives of the memorandum. In February 2015, in response to the OSTP memorandum, NIH released the *NIH Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research* (NIH Plan). The goals of the OSTP directive are in keeping with NIH's ongoing and future commitments to facilitate data sharing, and the NIH Plan outlines mechanisms for expanding and strengthening access to data and publications from NIH-funded research. Implementation of the NIH Plan will be consistent with the *Guiding Principles and Common Approach for Enhancing Public Access to the Results of Research Funded by HHS Operating Divisions*.

In an effort to move forward with its commitments to the data sharing enterprise and implement the NIH Plan, NIH is considering how to expand upon its 2003 Data Sharing Policy. For scientific digital data, NIH is considering requiring submission of data management plans by all NIH-funded research investigators, which will be evaluated during the peer review process. NIH also plans to encourage supported researchers to deposit data in established public repositories for archiving and preservation and to make use of existing data standards relevant to a specific research community, when applicable, to promote interoperability and downstream information processing. In order to ensure the discoverability of data sets resulting from NIH-funded research, NIH is considering developing a mechanism to index data sets, such as bioCADDIE, which will additionally facilitate the appropriate attribution to those responsible for the data, and link the data citations to associated publications. NIH is also considering developing a single shared space for basic and clinical research output, including data, software, and narrative. The NIH Data Commons is one space that would allow data from NIH-funded research to be available free of charge.

There are considerations and advice from the scientific community and from the public at-large that will help inform NIH in its further development of data sharing strategies and priorities. The course will focus on three of these strategies:

- Data Sharing Strategy Development
- Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications
- Long-term Sustainability and Value of Repositories

Data Sharing Strategy Development

Below are six axes related to the value of data for consideration:

1. **The purpose intended or expected to be served by sharing the data asset.** There are multiple potential purposes for data sharing, and those purposes may place different values on data elements, algorithms, software, and tools that are part of a data asset. These different purposes have different costs. For example, the cost of preparing the data to provide the reader of a scientific paper with a deeper understanding of how conclusions in that paper were reached could be lower than the costs of preparing data intended to be aggregated across different studies.

2. **Supporting data reuse and reproducibility of science.** Sharing of data elements, algorithms, tools and data necessary for addressing the primary aims/endpoints for which a data set was collected are critical to support FAIR. Secondary endpoints and the associated data elements, algorithms, tools and data should also be evaluated as additional targets for data sharing at a project level.

3. **The maturity of the science and the data infrastructure.** Some domains have a long tradition of data sharing and are supported by a robust infrastructure and culture. Data from these domains have consistent metadata, data description, and are machine readable. The most valuable of these data assets are well-annotated, including well-curated data, data elements, and metadata. Annotation using well-defined metadata resources, such as the Unified Medical Language System (UMLS) and other standard, open terminologies, vocabularies and ontologies, maximizes the value of such data assets and conforms to FAIR principles. Other domains, representing much of NIH-funded research, do not have such traditions nor infrastructures. For less data-centric research areas, the value of their data for sharing may depend more crucially on the purpose for which that sharing is intended (see Item 1).

4. **Uniqueness of the data.** Data may be more valuable if they are difficult to collect, such as data drawn from a rare or unique scientific circumstance, including access to rare biospecimens or populations exposed to a natural disaster.

5. **Urgency of the need for data.** Data to address an urgent health crisis, such as disease outbreaks, may be considered high value.

6. **Ethical considerations.** For example, the participation of human subjects in research is asserted by some to carry with it an ethical obligation to maximize the utility of the data through sharing, with appropriate attention to issues related to privacy and confidentiality.

Discussion Materials

1. Axes for assessing value and/or burden:

   a. The appropriateness of the six axes listed above for assessing the value and burden of sharing data.

   b. Additional axes or areas that should be considered in prioritizing data sharing activities.

   c. The relative value and priority of the axes (including any additional axes suggested).

2. Additional information requested:
   a. Domain(s) of research that are most important to you or your organization (e.g., cognitive neuroscience, infectious disease epidemiology)

   b. Data types generated or consumed by you or your organization (e.g., images, electronic health records, surveys)

   c. Repositories that are important to research of interest to you or your organization

d. The length of time data should be made available and the appropriate means for maintaining and sustaining such data

e. Data types that are not available but would be valuable for research of interest to you or your organization

f. Important purposes for reuse of data by you or your organization

g. Value of sharing unpublished data

h. Who can access the data

i. When data can be accessed (e.g. embargo periods, before publication)

j. Barriers or burden to data sharing

k. Barriers or burdens to data submission and deposit

l. Barriers and burdens in accessing data

m. Barriers or burden in reusing data

n. Burdens in data collection and maintenance

Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

Increased access to digital research data is expected to accelerate scientific discovery and progress. Effective data sharing relies upon identification, adoption, and crediting of good data management and sharing practices that support scientific rigor and reproducibility. Recognition of effective data sharing and reuse of high quality data can be supported through data citation.

To make publicly-funded data broadly accessible and to support reuse, reproducibility, and discovery, data often must be considered in conjunction with other related digital objects that make it understandable and reusable. These additional digital objects may include software tools, metadata, standards, or analytical workflows.

To be effectively shared, data should conform to the FAIR principles. Scholarly publications typically include citations to previously published research articles; these citations provide context for the motivation for the current study and the interpretation of the results presented in the publication. Many citations are limited to previous publications and the concepts within them, and do not cite the scientific data, software tools, or workflows that underlie them.

However, expectations of scholarly citation are evolving, and there is an apparent groundswell of support for data and software citation among the scientific research community. Data and software citation allows these important products of research programs be recognized and their impact and reuse assessed. Data citation allows researchers to have their high quality data sets attributed to them and may incentivize data sharing.

Data citations in publications can enable tracking and measuring impact and reuse of data sharing, and metrics for measuring data sharing, citation, and reuse may differ from those used in measuring impact of publications. As data sharing increases, development of informative metrics to ensure that data citation is effective and those who share highly reused data are acknowledged will become increasingly important. Some of the activities ongoing in this area include:

- Data citation is being actively pursued by the scientific community as a means to recognize the importance of FAIR data.

- FORCE11 group identifies eight principles that have been endorsed by over 100 organizations, including professional scientific societies, libraries, national standards bodies (e.g., National Information Standards Organization (NISO), educational institutions, funding organizations, libraries, data centers/repositories, and publishers).

- Publishers have begun to expect or require that data underlying publications are available with the publication and to include data citations, as well as to support freestanding data publications to further support data sharing.

- Clinical journal editors have called for the sharing of individual-level clinical trial data.

- FORCE11 has identified best practices for data citation with the research community, with the goal of allowing data citations to be included in the reference list of publications.

- The Earth Sciences Information Partnership (ESIP) has also developed extensive guidance for that research community to support consistent data citation.

NOTES

Current NIH Guidance on Research Project Progress Reports (RPPR) requires grantees to report other products of the research, which include data, databases, and software, in section C5a of their annual RPPR submission — yet there is little guidance on how to report them. The limited reporting of data and software sharing currently reported in the RPPR may reflect insufficient guidance from NIH on how to report data and software, or may be an indication of limited amounts of data and software sharing by the research community.

NIH recognizes that data and software citation provides proof of productivity above that provided by publications and patents. More thorough reporting of these in the RPPRs and in Competitive Renewals of Grant applications may strengthen documentation of productivity and also identify investigators and projects who most effectively share data and software. To that end, Data and Software citations should identify authors, and these authors may not be the same, or in the same order, as those for associated scientific publications.

Discussion Materials

1. The impact of data and software citation in incentivizing data sharing. NIH is interested in the impact of improved guidance on how to report on data and software resources that have been shared with the research community in research performance progress reports and competing grant applications.

   a. Impact of NIH providing improved guidance on how to cite and report data and software in the annual progress reports of NIH grants and research contracts.

   b. Consideration on peer review if guidance for competing renewals of grants strengthened reporting data and software sharing arising from the previous funding period.

2. When considering inclusion of data citations and publications in RPPRs or grant applications, NIH is interested in when researchers use a citation to data deposited into an existing, FAIR-compliant repository versus when they would submit a freestanding data publication.

3. Technical considerations of data citations and inclusion of other digital resources with data citations in NIH RPPRs and competitive Grant Renewals. NIH is interested in technical implementation of data citations, which may include:

   a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a DOI.

   b. Inclusion of a link to the data/software resource with the citation in the report.

   c. Identification of the authors of the Data/Software products.

   d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately.

   e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed.

4. NIH is interested in other routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals.

Long-term Sustainability and Value of Repositories

Colossal changes in biomedical research technologies and methods have shifted the bottleneck in scientific productivity from data production to data management, communication, and interpretation. Modern interdisciplinary team science requires an infrastructure and set of incentives to promote data sharing, and it needs an environment that fosters the development, dissemination, and effective use of computational tools for the analysis of datasets whose size and complexity have grown by orders of magnitude in recent years.

Digital data repositories represent a common mechanism for managing and storing biomedical content. The repositories enable specific communities to manage and preserve relevant data with the goal of ensuring continued existence and access to the data within the repository for the larger biomedical community. While there is a spectrum of models for content intake and management, biomedical digital data repositories can be thought of in two general categories: 1) deposition repositories, which

NOTES

support primary research data submitted by the data producers; and 2) knowledge bases, which provide curated findings derived from the aggregation or analysis of experimental data.

The increasing size and volume of biomedical data has led to increasing demand on biomedical data repositories. As research institutions begin to implement federal policies requiring them to share research data that have been gathered with the support of public funds, data repositories are growing in number, scale, and complexity. In this context, it is critical to understand and measure the value that these data repositories, and the individual data types and data sets that they contain, are providing to the research community. This information will support: 1) the ability of repository owners to prioritize activities related to the management of these repositories; 2) decisions by funding agencies which support biomedical data repositories; 3) communication about the usage and value of these repositories.

Discussion Materials

Qualitative and quantitative metrics such as those that describe:

- Utilization at multiple levels (repository, dataset, data item). In addition to the frequency of access and number of downloads, this might include:
    - Size and measured demand of the community served, placed in the context of the overall field.
    - The ongoing rate of data deposition and data access or download
- Indicators of data repository quality and impact. Examples include but are not limited to:
    - Publications from the data
    - Data citations
    - Altmetrics
    - Patents
    - Utilization of data sets in research studies
    - Outputs of those research studies, e.g. use in policies or guidelines
    - Enhanced data sharing and community collaboration around annotation/analysis of data sets
    - Economic measures such as investment and use value; efficiency impacts; return on investment
- Quality of service. Examples may include but are not limited to:
    - Implementation of a rigorous quality assurance process
    - Use of community-recognized standards
    - User support and training
    - Ease of data deposition and retrieval
    - Technical indicators, e.g., uptime, response time
- Infrastructure and governance. Examples may include but are not limited to:
    - Existence of an independent advisory board
    - Legal structure, e.g., access, security, licensing
    - Long-term sustainability plan

References

NIH Data Sharing Policies - http://grants.nih.gov/policy/sharing.htm

2003 Final NIH Statement on Sharing Research Data - http://grants.nih.gov/grants/policy/data_sharing

2014 NIH Genomic Data sharing Policy - https://gds.nih.gov/index.html

2015 NIH Intramural Human Data Sharing Policy - https://oir.nih.gov/sourcebook/intramural-program-oversight/intramural-data-sharing/human-data-sharing

2016 Policy on Dissemination of NIHI-Funded Clinical Trial Information - http://osp.od.nih.gov/office-clinical-research-and-bioethics-policy/clinical-research-policy/clinical-trials

White House Office of Science and Technology Policy: "Increasing Access to the Results of Federally Funded Scientific Research" - http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

2015 NIH Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research - http://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf

Guiding Principles and Common Approach for Enhancing Public Access to the Results of Research Funded by HHS Operating Divisions - http://www.hhs.gov/open/public-access-guiding-principles/index.html

bioCADDIE - https://biocaddie.org/

NIH Data Commons - https://datascience.nih.gov/commons

Sharing of Data Underling a Publication - http://www.hhs.gov/idealab/2015/02/27/hhs-expands-approach-making-research-results-freely-available-public

NL Yozwiak, et al., Nature 518, 477-479 (26 February 2015) doi: 10.1038/518477a - http://www.nature.com/news/data-sharing-make-outbreak-research-open-access-1.16966

S. Bull, et al., Journal of Empirical Research on Human Research Ethics 2015, Vol. 10(3) 225-238

FAIR Data:

SciData.2016;3:160018.doi:10.1038/sdata.2016.18 Joint Declaration of Data Citation Principles, FORCE11

Individual-level clinical trial data - http://www.icmje.org/news-and-editorials/M15-2928-PAP.pdf

Force 11 Best Practices for Data Citation - https://www.force11.org/

Earth Sciences Information Partnership (ESIP) - http://www.esipfed.org/

NIH Guidance on Research Project Reports (RPPR) - http://grants.nih.gov/grants/rppr/index.htm

Promoting Open Science: Data Sharing

Maryann E. Martone, PhD

Professor Emeritus, University of California, San Diego
Director of Biosciences, Hypothes.is
Founder, SciCrunch.com

NOTES

Almost all major funding agencies in the U.S. and abroad are pushing for more open science, which at a minimum encompasses the open sharing of the products of research including research articles, data and code (Gewin 2016; McKiernan et al. 2016). This push is driven by what I call the duality of modern scientific communications: our scientific results are no longer strictly consumed by other scientists. Rather, the consumers are both humans — scientists and non-scientists alike — and machines.

Prior to computers and the Internet, sharing research outputs routinely was not possible beyond what we could publish in books. Consequently, a culture grew up around scholarly publishing in which access was governed by physical constraints, data were considered disposable after some specified regulatory period, and the production of a data set on its own, in the absence of analysis was rarely considered a work of scholarship. The scientific article has held such a privileged place in scientific communication for so long that the current push to change questions its supremacy in the age of networks, big data, and machine learning. Prospective data sharing, where large data sets, e.g., the Human Genome, are commissioned and made available, is seen as a public good. But the routine sharing of relatively small data sets produced by small teams of scientists through often complex experimental designs, so called "long tail" data, has fewer advocates (Ferguson et al. 2014).

The opportunities afforded by computer algorithms to digest and synthesize the vast amount of data and prose produced by the biomedical community requires that we modernize our means of disseminating science for these new, disruptive technologies (the Internet, mobile devices). Few scientists fail to use software regularly in their interactions with research objects, defined here as narrative, data, and code. To the extent that we can make these research outputs free of any restrictions on their reuse, we enhance the ability for others to make use of research products, including combining data in new ways. A campaign is underway to make data and other research objects FAIR: findable, accessible, interoperable, and reusable for both humans and machines (Wilkinson et al. 2016).

The human side of open science emphasizes its benefit in addressing issues in reproducibility and transparency that are calling into question the integrity of scientific research (Gewin 2016; Open Science Collaboration et al. 2015). The human side also focuses on the ethical issues surrounding access by the public to research results largely generated through public funds.

To this end, we are seeing more calls to expand the notion of data sharing from controlled conditions established through peer to peer interactions, e.g., making data available on request, to a more open, e-science vision, where researchers conduct their research digitally and where data standards and programmatic interfaces make it easy for machines to access and reuse large amounts of data with the minimal amount of restriction possible.

Of course, where there is push there is pushback. (e.g., New England Journal of Medicine editorial on data sharing (Sharing 2016). In surveys and editorials across fields, the concerns and arguments against data sharing are remarkably similar:

1. "Professional vulnerability" (Rouder, 2015): Someone will use my data against me ("weaponizing data," "hostile replications") by finding errors in my work or otherwise deliberately trying to undermine my work by faulty replications.

2. Scooping: Someone (sometimes described as "research parasites") will do an analysis that I was planning to do, and claim the scientific credit for my work.

3. Time and effort: How will I be compensated for my time and other expenses for preparing the data for storage and retrieval/reuse?

4. Impenetrable data: My data are too complicated to understand and making them available may lead to bad science.

5. No one needs to understand them: I've already extracted all the value from the data and published them. I can't imagine how anyone would find use for them beyond what I have done.

6. Laziness and stagnation: No one will collect new data, just re-analyze the old.

In this session, we will consider the issue of data sharing from an ethical perspective. First, we will consider what is data? Is it a research asset, similar to the reagents and tools assembled by a laboratory for personal use? Is it an integral part of a scientific study that should be presented along with the introduction, methods, analyses and discussion? Is it a primary scientific outcome on equal footing with the narrative works we produce? Depending on how we view it, questions of how and when data should

NOTES

be shared have different answers. What does it mean to make data FAIR?

Next we'll consider who is helped or harmed by data sharing, in light of concerns about reproducibility and our views about what rights scientists have over their data. We will consider how current practices are changing in response to the need to make data available through development of new types of products designed around data, e.g, data journals, data papers and more formal systems for scientists to give and take credit for available data, e.g., data citation.

Finally, we'll consider some current case studies where scientists had both positive and negative experiences associated with replications of studies and sharing of data and tools. In the discussion period, we will consider what norms and "etiquette" could be developed to ease scientists' concerns about sharing their data, and whether current proposals help or hinder science.

Discussion Materials

1. A graduate student is given a project to reproduce findings in a prominent paper by a leading scientist published in *Cell*. The graduate student cannot reproduce the findings. The advisor feels that the graduate student is not doing a good job and communicates displeasure to the student. The graduate student gets discouraged and decides to drop out of the program; he meets with the head of the program who suggests he contacts the lead author of the study. When he does, he finds out that the published result only occurred in a subset of experiments — that is, there was a high failure rate within the original laboratory.

    a. What is the impact of such practices on science and its practitioners? Did the scientific process work in this case?

    b. "Whispers and innuendos" about reproducibility or lack thereof do little to advance science because they can't be acted upon by either party. How can science be both more rigorous and transparent in a way that respects both the process and the participants?

    c. How would this situation have played out if, as a condition of publication, all the data needed to be made available for inspection?

2. In light of "whispers and innuendos" in several communities about the lack of reproducibility, formal efforts have been made to reproduce key studies across different fields (Begley and Ellis 2012; Steward O, el tal., 2012; Open Science Collaboration et al. 2015). Efforts were made to contact and involve the authors of the original studies. Nevertheless, in all cases, the number of studies that could be reproduced was low. Published reports (e.g., Steward et al., 2011) clearly indicated the benefits of attempting these replications, but interviews with participants (e.g., Bohannon 2014) indicated that the attempts left some feeling bullied by the "self-appointed replication police." (Kahneman 2016) observes: "…The hypothesis that the failure is due to a flawed replication comes less readily to mind except for authors and their supporters, who often feel wronged." In light of the sensitivity involved in such studies, some have called for "replication etiquette" or norms to be developed around the handling of replications or the lack thereof.

    a. What is the proper balance between professional courtesy and the critical need for the "self-correcting process of science?"

    b. How should our current and future publication and reward system handle these cases?

3. We are still in the early days of grappling with research data and its place in the scientific process. In the arguments for and against, data sharing has been variously described as: 1) the detritus of science; 2) a supplement to the written record of science (Wallis et al., 2011); 3) a manifest of our intellectual output (Brembs, 2014); 4) a first class, independent research product.

    Consider these two cases:

    Case 3A: As a result of a clinical trial that examined the effects of antidepressants in adolescents, they are approved for use in children under the age of 18. The original study data were not made available, but in the wake of many adverse events, including increased suicide, calls were made through the "Restoring Invisible and Abandoned Trials (RIAT)" initiative for the data to be made available. A reanalysis conducted over a decade after the original report did not support the original findings of the study (Le Noury et al. 2015).

Case 3B: In response to reproducibility issues with translational research in spinal cord injury, the community made their individual data sets available to independent researchers, allowing aggregation across dozens of laboratories and thousands of animals with spinal cord injury (reviewed in Ferguson et al., 2014). Researchers shared not only their primary data, but animal care and laboratory records, so called "file drawer" and "background data". Reanalysis of this aggregate data set and similar efforts in TBI led to much better predictive models, pointed towards new therapeutic areas and provided more robust cross-species biomarkers of functional recovery. In this scenario, each laboratory contributed a slice of a large, multidimensional space represented by their individual experiments. It was sharing the data, including both positive and negative results, primary and background data, and not just hypotheses, protocols and results that led to a greater understanding of the phenomenon. Had this data sharing project not been attempted, all of this data would likely have been eventually lost.

a. How does one's views about sharing data and the norms, e.g., involving data providers in replications and embargo periods for data change depending on how data themselves are viewed? Is the concept of the "data parasite", i.e., one who lives off of other people's data, valid across all these contexts?

b. What changes need to be made in our publication and reward system to elevate the status of data?

References

Begley CG, Ellis LM (2012) Drug Development: Raise Standards for Preclinical cancer Research. *Nature* 483:7391, 531-533. Nature Publishing Group.

Bohannon J (2014) Psychology, Replication Effort Provokes Praise-and 'Bullying' Charges. *Science* 344:6186, 788-789.

Ferguson AR, Nielson JL, Cragin MH, Bandrowski AE, Martone ME (2014) Big Data from Small Data: Data-Sharing in the 'Long Tail' of Neuroscience. *Nature Neurosciencec* 17:11, 1442-1447.

Brembs B (2014) What is the Difference Between Text, Data and Code? Retrieved August 22, 2016 from http://bjoern.brembs.net/2014/03/what-is-the-difference-between-text-data-and-code/#anno.

Gewin V (2016) Data Sharing: An Open Mind on Open Data. *Nature* 529:7584, 117-119. Nature Publishing Group.

Garcia B (2016) Kahneman Commentary. Scribd. Retrieved August 26, 2016 from https://www.scribd.com/document/225285909/Kahneman-Commentary.

Hempel C, Oppenheim P, Meehl PE, Platt JR, Nosek BA, et al (2015) Psychology: Estimating the Reproducibility of Psychological Science. Science 349:6251. *American Association for the Advancement of Science.*

Le Noury J, Nardo JM, Healy D, Jureidini J, Raven M, Tufanaru C, Abi-Jaoude E (2015) Restoring Study 329: Efficacy and Harms of Paroxetine and Imipramine in Treatment of Major Depression in Adolescence. BMJ 351:7, h4320. British Medical Journal Publishing Group.

McKiernan EC, Bourne PE, Brown CT, Buck S, Kenall A, Lin J, McDougall D, et al (2016) How Open Science Helps researchers Succeed. eLife 5, 372-382. eLife Sciences Publications Limited.

Rouder JN (2015) The What, Why and, How of Born-Open Data. *Behavior Research Methods*, 1-8.

The International Consortium of Investigators for Fairness in Data Sharing (2016) Toward Fairness in Data Sharing. *The New England Journal of Medicine* 375:5, 405-407. Massachusetts Medical Society.

Steward O, Popovich PG, Dietrich WD, Kleitman N (2012) Replication and Reproducibility in Spinal Cord Injury Research. *Experimental Neurology*. doi:10.1016/j.expneurol.2011.06.017.

Wilkinson MD, Dumontier M, Aalbersberg IA, Appleton G, Axton M, Baak A, Blomberg N, et al (2016) The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3:March, 160018. Nature Publishing Group.