INFORMATICS PROFESSIONALS. LEADING THE WAY.

January 17, 2017

Carrie D. Wolinetz, Ph.D.
Associate Director for Science Policy
Office of Science Policy
National Institutes of Health

Submitted electronically at: http://osp.od.nih.gov/content/nih-request-information-strategies-nih-data-management-sharing-and-citation

Re: Request for Information: Strategies for NIH Data Management, Sharing and Citation

Dear Dr. Wolinetz:

The American Medical Informatics Association (AMIA) appreciates the opportunity to submit comments regarding the National Institutes of Health's (NIH) request for information (RFI) on data management, sharing and citation. AMIA is the professional home for more than 5,400 informatics professionals, representing researchers, front-line clinicians, public health experts, and educators who bring meaning to data, manage information and generate new knowledge across the health and research enterprise.

AMIA enthusiastically supports development of policies for data management, sharing and citation. Recently, AMIA published the first in a series of Policy Principles & Positions.[1] Among them was an articulation of our belief that data sharing among researchers is foundational to advance scientific discovery, foster a culture of transparency, and improve reproducibility. [2]

In considering this RFI, AMIA members identified three key institutional incentives as necessary to improve data management and sharing: (1) Making data sharing plans scorable elements of applicable grants; (2) Financially supporting data curation and sharing; and (3) identifying ways to support academic advancement for scholars who create and/or contribute to useful public datasets and software.

First, AMIA recommends NIH make Data Sharing Plans a "scorable" element of grant applications subject to the existing policy.[3] Data sharing has become such an important proximal output of research that we believe the relative value of a proposed project should include consideration of how its data will be shared. Making data sharing plans scorable would enable those

[1] "AMIA Public Policy Principles and Policy Positions, 2016 – 2017," available at http://bit.ly/2gPB52N

[2] Ibid., Data Sharing in Research Policy Principle (pg. 10)

[3] National Institutes of Health, "NIH Data Sharing Policy and Implementation Guidance," March 2003 https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

projects that prioritize systematic and strategic data sharing, through use of standards and accepted best-practice, to garner higher scores. By using the peer-review process, we will make incremental improvements to interoperability, while identifying approaches to better data sharing practices over time. Expert and peer review of data sharing plans will lead to improved data sharing across the NIH portfolio, which will greatly improve interoperability, and research rigor, transparency, traceability, and reproducibility. An important component of this recommendation is funding.

Second, AMIA recommends NIH earmark support for data sharing as part of applicable grants' direct costs. In order for researchers to dedicate additional time and energy to produce (or collaborate) on development and execution of a quality data sharing plan, specified funding is needed. Mandating robust sharing plans, and elevating them to be scorable without corresponding funding would be counterproductive, and likely diminish the impact of such a policy.

Finally, AMIA recommends NIH identify ways to provide institutional awards for scholars who create and/or contribute to useful public datasets and software. We note that our academic reward system does not adequately recognize or provide incentives for those who create as well as those who analyze data.[4] NIH should look to scale platforms such as dbGaP, which enable investigators to receive points for others using their data, to other types of research.

Below we outline our recommendations in more detail, and we address NIH's specific questions related to this RFI. A group of AMIA members, listed in Appendix A, has provide detailed responses to this RFI in Table 1 of the enclosed document. These responses have been reviewed and duly approved by AMIA's Public Policy Committee and the AMIA Board of Directors. Should you have any questions or require additional information, please contact AMIA Vice President for Public Policy Jeffery Smith at jsmith@amia.org or (301) 657-1291 ext. 113. We, again, thank NIH for the opportunity to comment and look forward to continued dialogue.

Sincerely,

Douglas B. Fridsma, MD, PhD, FACP, FACMI
President and CEO
AMIA

Thomas H. Payne, MD, FACP, FACMI
AMIA Board Chair
Medical Director, IT Services, UW Medicine
University of Washington

Enclosed: Detailed AMIA Recommendations and Comments to NIH Questions

[4] Piwowar, H., Vision, T., "Data reuse and the open data citation advantage," *Peer J.* 2013. 1:e175

Detailed AMIA Recommendations and Comments to NIH Questions

SECTION 1. Data Sharing Strategy Development

High-priority types of data to be shared

Any data necessary to the process of reproducing research are of high-value. While we understand this casts a wide net, we believe reproducibility and provenance are especially important in the basic sciences where investigators are using pre-clinical or other data to achieve the same results, in large epidemiological data sets used for population-based research, and data used in the review of pharmaceutical and medical devices.

We also note sharing is most important for data that would be expensive to re-create. However, if the data are inexpensive to generate, clear methodological instructions might be sufficient for re-creation, making sharing less important.

The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications

In discussing length of time, AMIA members note a common tension concerns how quickly data should be shared, in addition to how long they should be public. We applaud NIH for finalizing policy, concerning how quickly data must be deposited in ClinicalTrails.gov,[5] and encourage it to harmonize similar requirements beyond the registry and results database.

AMIA recommends NIH endeavor to make publicly-funded research data available for secondary use for at least ten years after it is first published, and we recommend NIH develop policies for both unrestricted and controlled access. Data governance and permission policies should be an additional area of NIH focus. We note the NIMH Limited Access Datasets (LAD) project as an exemplar of defining access requirements. In whatever way NIH proposes to identify data retention and access policies, AMIA recommends NIH harmonize requirements across its Centers and Institutes.

Where possible, NIH should leverage existing and proven environments to maintain and sustain publicly-funded data through platforms such as Dryad,[6] Dataverse,[7] Cancer Imaging Data,[8] Figshare,[9] Zendo[10] and BioCADDIE.[11] Consistent with previous AMIA recommendations on

[5] 42 CFR Part 11. Clinical Trials Registration and Results Information Submission; Final Rule.
[6] http://datadryad.org
[7] https://dataverse.harvard.edu
[8] http://www.cancerimagingarchive.net
[9] https://figshare.com
[10] https://www.zenodo.org
[11] https://biocaddie.org

digital data repositories, we support development of metrics to evaluate the quality and fit-for-purpose of various repositories.[12] Key among these metrics should be consideration of the repositories sustainability and/or business model. Should NIH designate an existing, independently operated repository, researchers depositing data need to be assured of their continued existence and availability. Additionally, NIH should prefer repositories that store the data in a non-proprietary (i.e. open) data format. Should a repository shutter or fail to meet its contractual obligations, it is important to protect the data from being "locked-in."

Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers

AMIA supports dedicated funding from research sponsors for data curation and donation efforts so there are sufficient incentives to share, collaborate, and advance data sharing capabilities.[13] We recommend NIH earmark a percentage of grant funds for such activities as a way to overcome cost barriers. In combination with scoring data sharing plans (DSPs), explicitly setting aside funds to carry-out the DSP will improve data stewardship and sharing. Further, ensuring adherence to FAIR principles – Findability, Accessibility, Interoperability and Reusability – will help demonstrate value to overcome cost concerns.

Any other topics respondents recognize as important for NIH to consider

Additional aspects NIH may want to consider surround the role of participants in research data sharing. Specifically, NIH should ensure that data sharing policies are clearly articulated to both researchers and patients; that there are mechanisms for consent management; provisions of notification when data is used, and ways to share / return results in appropriate circumstances.

Additionally, NIH may wish to articulate expectations around pre-publication data management, including annotation, metadata, and provenance. For example, NIH should consider what documentation should be shared along with data that are necessary to support reuse, such as processes for transformation, imputation, coding, mapping standardization, data cleaning, and data quality assessments.

Finally, NIH should develop guidelines and best practices around data discoverability, including the use of model annotations, metadata schemas focused on a given domain (e.g. imaging) and minimal metadata expectations.

[12] AMIA Response to NIH RFI on Metrics to Assess Value of Biomedical Digital Repositories, October 5, 2016 available at http://bit.ly/2i2XF5a

[13] Borne, P., Lorsch, J., Green, E., "Perspective: Sustaining the big-data ecosystem," *Nature*. November 2015. 527, S16–S17

Section II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports (RPPR) and Grant Applications

The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing

We see the utility in this kind of requirement; however, we reiterate our recommendation that there exist dedicated funding from research sponsors for data curation and donation efforts. Additionally, we recommend that reporting requirements be shared across venues (i.e. RPPR, publication in journals, etc.) with common guidance and metadata wherever possible. We further note that multiple approaches to point towards data, such as through data repository URLs, software source code hosting services, and DOIs, should be supported, and granularity issues should be supported in metadata models whenever possible.

Important features of technical guidance for data and software citation in reports to NIH, which may include:

- *Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI) (https://www.iso.org/obp/ui/#iso:std:iso:26324:ed-1:v1:en)*
- *Inclusion of a link to the data/software resource with the citation in the report*
- *Identification of the authors of the data/software products*
- *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*
- *Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed;*

AMIA strongly supports development of policy to cite data and software developed by grantees. As discussed previously, institutional incentives are needed to encourage development of reusable data / software. Data journals[14] are a step in the right direction, and development of persistent UIDs, such as DOIs, for data / software citation would be an important contribution to this effort. However, we note that such DOIs for re-use are new and nascent. We encourage NIH to fund specific projects to improve the use of DOIs for data / software, and we encourage NIH to explore how open source code and software containers, which represent snapshots of entire operating system configurations of computers used to develop software, can be leveraged to improve research rigor.

Inclusion of data sharing activities in scientists' career assessments is a potentially powerful means of incentivizing data sharing. Recent "Alt-metrics" efforts have begun to build the framework for

[14] http://www.nature.com/sdata/

tracking data reuse and citation as meaningful measurements of researcher contributions. As a means to help develop these policies and to further encourage data sharing, AMIA recommends NIH host a roundtable of academic medical leaders, and produce a handbook for integrating this type of "credit" into promotion and tenure decisions.

Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications;

Again, making data sharing plans scorable aspects of pertinent grant applications would enable reviewers to assess the mechanisms through which data / software will be shared, and it would encourage more systematic, robust sharing strategies. One example for how NIH could operationalize the scoring of data sharing plans, could be to score according to priority data types (e.g. sharing data for autism or rare disease) and data quality / usability, similar to the 5-star deployment scheme for Open Data.[15]

Further, development of policy around Data Management Plans (DMPs), and examples of DMPs, would strengthen data and software sharing, especially towards to goals of traceability and reproducibility of research. According to research in progress, shared for purposes of AMIA's response to this RFI, a review of 67 data management and sharing plan requirements documents "uncovered inconsistent requirements for written DMPs as well as high variability in required or suggested DMP topics among funder requirements."[16] Further, "DMP requirements were found to emphasize post-publication data sharing rather than upstream activities that impact data quality, provide traceability or support reproducibility. With the emphasis equalized, the forty-three identified topics can aid Data Managers in systematically generating comprehensive DMPs that support project planning and application evaluation as well as data management conduct and post-publication data sharing."

Any other topics respondents recognize as important for NIH to consider.

We point NIH reviewers to real-time open science projects that have developed citable reference, which add to the diversity of solutions available. Through a project called Thinklab, researchers can get feedback on their grant proposals, participate in open peer review, and even lead entirely open research projects.[17] An example of an open science project developed using Thinklab is "Rephetio: Repurposing drugs on a hetnet," where each of the discussions that were part of that project is developed as a citable reference.[18,19] For example, https://thinklab.com/discussion/incorporating-

[15] http://5stardata.info/en/
[16] Zousus, M., Williams, M. "Data Management Plans, The Missing Perspective," available in a forthcoming issue of *Journal of Biomedical Informatics*
[17] https://thinklab.com/about
[18] https://thinklab.com/p/rephetio
[19] https://thinklab.com/p/rephetio/discussion

drugcentral-data-in-our-network/186 can be cited as "Daniel Himmelstein, Oleg Ursu, Mike Gilson, Pouya Khankhanian, Tudor Oprea (2016) Incorporating DrugCentral data in our network. *Thinklab*. doi:10.15363/thinklab.d186." While we acknowledge that such citations do not have the same rigor as peer-reviewed publication it may provide another avenue for recognition for researchers.

We also point NIH reviewers to the concept of "nanopublications" as means to disseminate individual data as independent publications with or without an accompanying research article.[20] Because nanopublications can be attributed and cited, they provide incentives for researchers to make their data available in standard formats that drive data accessibility and interoperability.

Increasingly, the research community is becoming aware that reproducibility requires that all software used to collect, transform and analyze the data must be publically available for inspection, modification, and reuse, along with the data.[21] We encourage NIH reviewers to become more sensitive to open source concepts relevant to ensuring scientific software practices support reproducibility such as the value of (1) using well-defined, standard open source licenses approved by the Open Source Initiative (OSI)[22] and (2) open source best practices of a public code repository (e.g., on GitHub[23] or Bitbucket[24]) and a public issue tracker.

Finally, any guidelines or requirements issued by NIH should consider *when* data must be shared. Timing requirements for data sharing may require tradeoffs between the goals of data originators hoping to retain exclusive access to data as needed to publish papers and those interested in prompt access to data for secondary use and replication. The NIH should develop models and policies designed to balance these potentially competing needs. Models such as pre-publication of protocols, as required by ClinicalTrials.gov, might be a partial solution. Alternatively, timeliness of data sharing efforts might be considered in the review of data management plans.

[20] http://nanopub.org/wordpress/
[21] Ince, D., et al., "The case for open computer programs." *Nature. February 2012.* **482**, S485-488. doi:10.1038/nature10836
[22] https://opensource.org/licenses
[23] https://github.com/
[24] https://bitbucket.org/

Appendix A: Response Team Members

These comments are official AMIA policy and endorsed by the full AMIA membership and Board of Directors, as evidenced by the President & CEO and Board Chair signatures. Below are the names of AMIA members who participated in the drafting and development of these policies recommendations.

Response Team: NIH RFI on Data Management, Sharing & Citation			
First	**Last**	**Organization**	**Email**
Tudor	Oprea	UNM School of Medicine	toprea@salud.unm.edu
Edwin	Young	Mount Sinai Health System	edwin.young@mountsinai.org
Michael	Cantor	NYULMC	michael.cantor@nyumc.org
Jorge	Caballero	Distal, Inc.	jorge@distal.co
Satyajeet	Raje	National Library of Medicine	satyajeet.raje@nih.gov
Harry	Hochheiser	U. Pittsbugh	harryh@pitt.edu
Ashish	Sharma	Biomedical Informatics, Emory Univ.	ashish.sharma@emory.edu
Rashmi	Mishra	NIDCR	rashmi.mishra2@nih.gov
Eileen	Healy	NIH	healye@slu.edu
Brian	Fung	Mayo Clinic	fung.brian@mayo.edu
Tod	Yates	Blue Cross Blue Shield AZ Advantage	tod.yates@azbluemedicare.com
Meredith	Zozus	Univ. of Arkansas for Medical Sciences	mzozus@uams.edu
Ken	Goodman	University of Miami	kgoodman@med.miami.edu
Leon	Rozenblit	Prometheus Research, LLC	leon@prometheusresearch.com
Jorge	Ferrer	VA	Jorge.Ferrer@va.gov
Carolyn	Petersen	Mayo Clinic	Petersen.Carolyn@mayo.edu