FASEB
Federation of American Societies for Experimental Biology

Representing Over 130,000 Researchers

301.634.7000
www.faseb.org

9650 Rockville Pike
Bethesda, MD 20814

November 8, 2017

The Federation of American Societies for Experimental Biology (FASEB) appreciates the opportunity to provide feedback to the Request for Information (RFI) from the National Library of Medicine (NLM) entitled "Next-Generation Data Science Challenges in Health and Biomedicine" (NOT-LM-17-006). Comprising 31 scientific societies, which collectively represent over 130,000 biological and biomedical researchers, FASEB supports NLM's goal to strengthen and expand the scope of biomedical data science research. The Federation commends NLM for engaging the community with this RFI, and hopes it will continue to seek stakeholder input as it finalizes and implements its strategic plan. Below is FASEB's response to this RFI, submitted online at https://www.research.net/r/NLMDataSci.

1) Promising directions for new data science research in the context of health and biomedicine.

Data science research has the potential to transform all fields of the biomedical and health sciences. Through PubMed and associated platforms (PubMed Central and PubMed Commons), NLM pioneered new ways in which researchers find, access, use, and cite scientific manuscripts. FASEB hopes that NLM's data science initiatives will lead to similar innovations for datasets, databases, analysis tools, and other core data resources. We also urge NLM to coordinate ongoing and future efforts across NIH to ensure integration, interoperability, and the adoption of data management best practices.

Development of new core data resources

Investigators rely upon and require more core data resources. In FASEB's recent survey of shared research resources, the majority of respondents reported that specialized software, advanced IT infrastructure, and databases/knowledgebases were essential to their work. However, when asked about significant unmet needs, participants frequently selected these same three categories (see Questions 15 and 17 in Appendix A). New research initiatives, such as the Human Cell Atlas and Mouse Brain Reconstruction Project, are producing a plethora of large and complex datasets. To effectively utilize this information, investigators need innovative data analysis tools. Similarly, novel analytics will be required to take full advantage of rapidly evolving imaging technologies. Each of these needs affects multiple scientific domains.

Sustainability of existing core data resources

The long-term sustainability of most databases and other core data resources is uncertain. These resources are crucial to advancing the biological and health sciences. For example, topical databases make data discoverable and accessible, as well as play an essential role in establishing and promulgating nomenclature and data standards within their fields.

FASEB recommends that NLM support initiatives to improve and implement standard business models and infrastructure for these resources. This will require a multi-pronged approach, including the creation

of novel curation tools to reduce operational costs and development of frameworks to determine which datasets should be preserved and for how long. Sustainable business models must also establish clear expectations for all stakeholders and define responsibilities for maintenance throughout the data lifecycle.

Data hosting and access entail costs to the provider. To mitigate the impact of these expenses on research budgets, NIH has sought to combine some databases – such as the most recent effort to merge the model organism databases. Combining databases with different nomenclatures, organization, and documentation of data is a major undertaking. FASEB is concerned that there is insufficient funding and technical expertise available to accomplish this ambitious goal. NLM should ensure such efforts receive sufficient financial and technical support and seek "lessons learned" to inform other database initiatives.

2) Promising directions for new initiatives relating to open science and research reproducibility.

The data sciences can enhance open science and research reproducibility. Core data resources, such as databases and knowledgebases, allow investigators to replicate prior analyses and reuse data in new research projects. Analytics and other data-related tools further enable reuse.

Coordination of data sciences initiatives

FASEB recognizes NLM's strength and expertise in the data sciences, and we are pleased to see that NIH's data science efforts have found an established home at NLM. However, FASEB is concerned that NLM remains functionally isolated from many NIH institutes and centers (I/Cs). Without NIH-wide coordination, individual I/Cs may develop data resources that are not fully interoperable with those created by other I/Cs. For example, several I/Cs are developing separate, independent clouds for their core data resources. This fragmented approach could limit the ways data are accessed, reused, and combined – an outcome inconsistent with the goals of open science. In addition, this fragmented strategy can produce needless redundancies, thus increasing the cost of research. NLM will need to continue to develop partnerships across the individual NIH I/Cs and promote greater coordination of core data resources and initiatives.

Development and deployment of metadata standards

Research reproducibility depends upon rigorous experimental design and interpretation and application of resulting data. Metadata – or description of a dataset – provide essential information for determining appropriate use. Robust, well-accepted metadata standards do not exist for many fields or many data types. Furthermore, minimal metadata standards have not been established or deployed across all NIH databases. Therefore, FASEB calls upon NLM to support the development of community-based metadata standards. Scientific societies can aid these efforts by helping to identify and convene subject matter experts and disseminating consensus standards.

We also urge NLM to lead the development of automated tools for assigning metadata to files and datasets. Development of these tools can begin before census standards are established. Through automation, investigators will not have to personally keep track of metadata standards; the tools can be updated instead. Thus, metadata tools could speed adoption of new standards and changes to existing standards while also minimizing burden on the research community.

Creation of tools that increase interoperability and facilitate good data practices

FASEB recommends that NLM support the development of tools that increase the interoperability of data at every stage of the data lifecycle – from collection to aggregation, deposition, access, and reuse. For example, FASEB and other stakeholders have called for the creation of import/export tools that can interface between ClinicalTrials.gov and clinical trial management software. Such tools would reduce the human effort required for data sharing and prevent errors introduced during conversion or manual entry. To the greatest extent possible, interoperability tools should automatically extract or deduce metadata.

Data citation promotes reproducibility by clarifying exactly what data were utilized in a study, and new tools can enable this practice. Currently, however, citation of research articles is simpler and more expedient than citation of datasets. Commonly used citation management platforms offer limited-to-no support for data citation. Similarly, many databases, including those managed by NIH, do not offer a citation export function, like the one available for articles indexed in PubMed. To promote greater data citation, FASEB recommends that NLM's initiatives address both of these gaps.

3) Promising directions for workforce development and new partnerships.

Data literacy across the scientific workforce

NLM's workforce development efforts should take into consideration researchers that are not data science specialists. Ideally, all investigators should be able to handle basic data-related tasks and identify when they need to collaborate with a data scientist, statistician, or informaticist. However, many investigators have not received formal training in data management. Furthermore, the data sciences are evolving so quickly prior training may quickly become outdated. This situation makes it challenging for individual researchers and entire fields to achieve good data practices. NLM should support training opportunities and self-paced educational resources for the entire research workforce. In addition, NLM should explore ways to alert research communities of major advancements in the data sciences and the availability of new tools that enable good practices.

4) Respondents may also propose additional ideas related to health and biomedicine

Evaluation of BD2K

Significant funding and effort has been expended in support of the Big Data to Knowledge (BD2K) initiative, with the goal of enhancing data science across the biological research enterprise. However, NIH

has not publicly released any comprehensive assessments of this program. To inform discussions of new initiatives, such as in response to this Request for Information (RFI), stakeholders require more information about this program's outcomes to date. FASEB recommends that an evaluation of BD2K consider the following questions: Is BD2K achieving its original goals? To what extent have products or findings been adopted or implemented by the research community? What impact has it had on the broader research workforce? What important data science issues or areas were not addressed by BD2K (i.e., what gaps exist)?

Greater engagement and responsiveness

Biological researchers are not merely users of the fruits of NLM's labors but often the creative force that generates the data, develops the computational methods and software for facile analysis, and establishes new databases that then become the community standards that NLM magnifies and distributes. Thus, we urge the NLM to maintain and enhance responsiveness to the scientific community regarding the development and maintenance of data, computational methods, and database resources.

Use of Categories in this Request

The "Data-driven Discovery" versus "Data-driven Health Improvement" categorization scheme used in this RFI presents a false dichotomy. Research takes place across a spectrum and can lead to discoveries in different areas than originally anticipated. Furthermore, many researchers draw upon basic and clinical data in their work, and many datasets can be used for both "discovery" and "health improvement" research. Therefore, FASEB encourages NLM to consider using distinctions based on the type of user (researcher, clinician, patient, or member of the public) rather than the anticipated research outcome.