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

April 4, 2018

Comments submitted via email: DataScienceRFI@mail.nih.gov

Dear NIH Scientific Data Council and NIH Data Science Policy Council,

The Federation of American Societies for Experimental Biology (FASEB) thanks the National Institutes of Health (NIH) for the opportunity to provide feedback on NIH's Strategic Plan for Data Science (NOT-OD-18-134). FASEB is comprised of 30 member societies, collectively representing over 130,000 biological and biomedical researchers who produce and use a wide variety of data, core data resources, and analytics.

The five goals articulated in the proposed strategic plan reflect key challenges to modernizing the data resource ecosystem. FASEB offers the following four cross-cutting recommendations as well as specific feedback on each strategic objective. We hope these comments will aid NIH as it finalizes this plan and encourage NIH to provide additional opportunities to offer feedback during its implementation and assessment.

Cross-cutting recommendations:

Rapidly deploy at least minimal data standards: Data standards are necessary to realize NIH's goal of ensuring adherence to the FAIR principles; without standards, much biomedical data cannot be reused or even reassessed. In healthcare, lack of standardization of health record systems has led to siloed data, hindering analyses across systems and reducing portability of medical records. The community of data producers and consumers is already working towards standards development, and we urge NIH to consider partnership with them to avoid duplication and facilitate adoption. FASEB also encourages NIH to rapidly establish minimal data standards across all its databases and resources. To build upon the minimal data standards, FASEB recommends that NIH support community-based development of standards for specific data types and fields of research.

Consider the long-tail of biological research data: Many NIH-supported investigators produce and use datasets from individual, small-scale studies, also known as long-tail data. Much of this long-tail data does not fall within the scope of discipline-based or government repositories. NIH should be cognizant of biological research's long-tail as it determines the rate and progress of implementation. In particular, NIH's development of new policies and grant requirements must take into account these researcher's resource and accessibility needs as well as their uneven technical capabilities in the field of data science. Furthermore, the cost versus utility of making some long-tail data FAIR may not be justifiable, and implementation plans should take this constraint into account. NIH also must address incentive structures to ensure that these "long-tail" investigators will participate in and benefit from the emerging data ecosystem.

Ensure NIH-wide coordination, participation, and leadership: To achieve the goals outlined in its strategic plan, NIH must secure participation and coordination across all of its 27 institutes and centers (I/Cs) – a challenge we raised in our recent comments to the National Library of Medicine (NLM). If there is no trans-NIH coordination in the development and support of core data resources, the resulting efforts and products will lack interoperability. However, the plan does not describe how NIH-wide coordination will be accomplished or define the roles of the new Chief Data Strategist, the Office of the Director, or NLM. Therefore, FASEB recommends that NIH clarify who will be responsible for overseeing implementation of the strategic plan and ensure that they are able to carry out this important and complex initiative.

Involve FASEB and other stakeholders: The proposed strategic plan will affect a wide-range of stakeholders. To ensure integration of all perspectives in the implementation plan, FASEB strongly recommends that NIH engage and collaborate with data producer, consumer, and repository communities. Involving the broader scientific community in the plan's implementation and assessments will ensure relevance and facilitate adoption. Stakeholder engagement also could help NIH identify cost-efficient implementation strategies, ascertain and mitigate burden for researchers, and verify that investigators are receiving sufficient support to achieve the FAIR principles.

Feedback on strategic objectives:

Goal 1: Support a Highly Efficient and Effective Biomedical Research Data Infrastructure

Objective 1-1: Optimize data storage and security

By creating an online environment to facilitate access to large, high-value datasets, NIH can help ensure these datasets achieve the FAIR principles. **When working with cloud providers, FASEB encourages NIH to develop strategies to ensure that these services are available and affordable for individual investigators.** We also recognize that NIH may be better served in some cases by using non-cloud, internal servers to handle baseline computing needs and employing cloud services to cover computing surges and outages. The NSF High Performance Computing Centers are an example of alternatives to commercial cloud providers. Regardless of approach, sustainability, accessibility, and portability across platforms must be key factors in these decisions.

Objective 1-2: Connect NIH data systems

Linking valuable datasets facilitates reuse and enables more lines of inquiry. FASEB strongly supports NIH's plan to link the NIH Data Commons and widely-used NIH databases, as well as develop connections with non-NIH data resources. By breaking down these silos, NIH will provide a roadmap to the research community on how to integrate disparate databases.

Goal 2: Promote Modernization of the Data-Resources Ecosystem

Objective 2-1: Modernize the data repository ecosystem

FASEB agrees with NIH's assessment that the grant programs used for research projects are ill-suited to support and evaluate core data resources. NIH's intention to create a comprehensive funding mechanism and review criteria for core data resources is laudatory and will enable appropriate and robust peer review of these grant applications. However, the plan's definitions for core data resources and proposed separate funding mechanisms for resources and tools need additional consideration to ensure development of a highly integrated and sustainable data infrastructure.

FASEB is particularly concerned about NIH's intention to establish separate funding programs for databases and knowledgebases. In many cases that distinction is subjective; data science experts may be unable to agree whether a given type of data belongs in a database or a knowledgebase. This distinction may also penalize valuable resources that maintain and link a mixture of "core data" and "information related to core data" (another distinction we find to be ambiguous), when, in fact, integration of multiple data types within a "hybrid database/knowledgebase" is highly beneficial to end users. **Therefore, FASEB recommends that NIH proceed with establishing a single funding mechanism for investigator-driven development and support of databases, knowledgebases, and mixtures of the two.**

Similarly, NIH's intention not to fund any tool development as part of database grants is problematic. Many tools are integral to basic database development and functions (such as curation, query, data validation, and web presentation software); thus, separating tool development from a database project would reduce the utility of the resulting database. Instead, FASEB recommends that NIH use the grant review process to determine on a case-by-case basis whether proposed tool development in an application is justified and not duplicative; if not, it should be excised from the application.

Objective 2-2: Support the storage and sharing of individual datasets

Providing investigators with solutions for the storage and sharing of datasets will remove a key barrier to data accessibility. NIH's plan to eventually accept submission of individual, FAIR datasets into the Data Commons is commendable. However, this new environment will require a new set of incentives to encourage participation. **FASEB recommends establishment of an incentive structure to encourage appropriate data deposition and citation.**

Objective 2-3: Leverage ongoing initiatives to better integrate clinical and observational data into biomedical data science

To protect participant confidentiality and safety, access to clinical data is understandably limited. However, obtaining access to datasets stored in different repositories can be a daunting task for investigators, and it can be difficult to identify and correct any duplicative entries between datasets.

By creating linkages among NIH data resources, instituting universal credentialing protocols and authorization systems, and promoting use of the NIH Common Data Element Repository – as detailed in its strategic plan – NIH can reduce the administrative burden and technical challenges to appropriate and ethical use of this data.

Goal 3: Support the Development and Dissemination of Advanced Data Management, Analytics, and Visualization Tools

The emergence of "big data" is allowing investigators to pursue more lines of inquiry that could ultimately lead to transformative discoveries. However, as larger quantities and more types of data can be combined in new ways, we must also be cautious of spurious correlations and "over-mining" of datasets. The Federation is concerned that analytical methods and tools do not always keep pace with research opportunities. Rigorous research practices will depend on the continued development of analytical methods, which requires significant investment. **Therefore, FASEB recommends the addition of an objective under Goal 3 that provides support for research on "big data" analytical methods and best practices**.

Objective 3-1: Support useful, generalizable, and accessible tools and workflows

As noted in our comments on Objective 2-1, separating funding mechanisms for data resources from research projects is appropriate. A specialized program for tool development will allow NIH to establish appropriate metrics and review panels for these applications. However, we once again oppose any blanket refusal to consider funding tool development as part of database and knowledgebase projects.

FASEB encourages NIH to look towards successful projects from BD2K, international research efforts, and other U.S. research agencies for additional tool prototypes. However, we caution against dependence on commercial products that are not open source, which could limit future development and application of NIH-supported tools.

Objective 3-2: Broaden utility, usability, and accessibility of specialized tools

Specialized tools developed in one field can be repurposed by researchers in other fields, providing novel sources of tools and analytical methods for biomedical research. By encouraging developers to use container technologies such as Docker, NIH can promote greater access to and adoption of NIH-supported tools. As with other digital objects, shared tools should be tagged with a unique identifier to allow producers to receive credit for their contributions.

Objective 3-3: Improve discovery and cataloging of resources

Data citation helps make datasets findable and accessible and incentivizes data sharing. However, researchers often cite the corresponding article instead of the dataset because it is simpler and more expedient. To promote appropriate data citation, **FASEB recommends the following addition to the implementation tactics for Objective 3-3: "add exportable citation information to all NIH**

database entries, similar to what is provided for articles indexed in PubMed." These export features should be compatible with citation management software.

Goal 4: Enhance Workforce Development for Biomedical Data Science

Objective 4-1: Enhance the NIH data-science workforce

The proposed NIH Data Fellows program has the potential to incorporate expertise from diverse data science fields into ongoing biomedical research projects. If successful, **FASEB encourages NIH to expand the scope of this program to major extramural projects** that could similarly benefit from cutting-edge skills in the data sciences.

Objective 4-2: Expand the national research workforce

In prior statements,[1] FASEB detailed the need to increase data literacy across the entire scientific workforce. We are concerned, therefore, that the training activities proposed under this objective are limited to the graduate and postdoctoral career stages. Investigators, staff scientists, and technicians also should have access to educational resources that could strengthen their research projects. **FASEB recommends that NIH support learning opportunities for individuals at all career stages.** This could take the form of special sessions or workshops at scientific conferences, MOOCs and online modules, and consultations with subject matter experts. We also suggest making the data science training programs for NIH staff, described in Objective 4-1, available to all NIH grantees.

Objective 4-3: Engage a broader community

When making biomedical research data accessible, it is important to consider how the broader community will use the shared data, particularly when applied to medical decisions. In comments regarding dbGaP, FASEB recommended development of an educational module designed for healthcare providers. We are pleased to see that the NIH Strategic Plan for Data Science includes the creation of training materials for clinicians.

Goal 5: Enact Appropriate Policies to Promote Stewardship and Sustainability

Objective 5-1: Develop policies for a FAIR data ecosystem

FASEB commends NIH for affirming that robust data ecosystem policies must be "achievable and sustainable, and do not impose unnecessary burdens or untenable expectations on grantee institutions." We also recognize that sharing and access requirements that extend beyond the life of a grant can create an unfunded mandate. Any NIH policies to make resulting data FAIR must be accompanied by modifications to grant programs to defray the associated costs.

---

[1] For recent examples, please refer to the following: FASEB comments on NLM RFI regarding new data science opportunities (November 8, 2017); FASEB comments on NIH RFI, Processes for dbGaP Data Submission Access and Management (April 5, 2017); and FASEB response NIH RFI on strategies for data management sharing and citation (December 7, 2016).

FASEB also encourages NIH to harmonize its policies with other federal research agencies. For example, most agencies have adopted the use of data management plans (DMPs) for all research grant applications. FASEB strongly supports the use of DMPs because they: (1) encourage researchers to consider how they will collect, process, store, and share data; (2) can be tailored to a research project and available resources; (3) clarify expectations between grantees and sponsors; and (4) help sponsors identify infrastructure and workforce gaps. The FASEB's Statement on Data Management and Access offers recommendations for DMP requirements and compliance reporting.

Objective 5-2: Enhance stewardship

In previous comments to NLM, FASEB raised the issue of core data resource sustainability. Therefore, we are appreciative of NIH's emphasis on the need for stewardship and long-term stability of these resources in its Strategic Plan for Data Science. However, it will be difficult to develop business models that are appropriate for the disparate types of data resources. As a first step, FASEB suggests that NIH consider strategies to reduce costs and develop a long-term vision for the role of data science in biomedical research.

Expensive data management processes, as noted in the plan's introduction, could erode support for generating new data. Automating steps in data creation, sharing, and reuse can promote financial stewardship and good data practices. **Therefore, FASEB recommends that NIH fund research and tool development that will automate many steps of the data lifecycle.** Costs can be further reduced by allowing data to be deleted if they are no longer timely or useful or, at least, placed into some form of "cold storage" that does not impose a significant maintenance burden.

Finally, good stewardship requires a long-term plan; however, the strategic plan as written offers little insight into how revolutions in data science could shape the way biomedical research is conducted. Therefore, FASEB urges **NIH to consider what the research laboratory of the future will require as it determines what infrastructure, tools, and workforce development to support today**.

FASEB thanks NIH for engaging stakeholders in the finalization of its Strategic Plan for Data Science. In light of the tremendous opportunities data science offers biomedical research, it is imperative that NIH and the research community work together to leverage these advances. Please do not hesitate to contact me if FASEB can provide further assistance.

Sincerely,

Thomas O. Baldwin, PhD
FASEB President