April 2, 2018

National Institutes of Health
9000 Rockville Pike
Bethesda, MD 20892

Re: NOT-OD-18-134: Request for Information (RFI): Soliciting Input for the National Institutes of Health (NIH) Strategic Plan for Data Science

Submitted via http://grants.nih.gov/grants/rfi/rfi.cfm?ID=73

Dear Sir/Madam:

Biocom is the largest, most experienced leader and advocate for California's life science sector, which includes biotechnology, pharmaceutical, medical device, genomics and diagnostics companies of all sizes, as well as research universities and institutes, clinical research organizations, investors and service providers. With more than 1,000 members dedicated to improving health and quality of life, Biocom drives public policy initiatives to positively influence the state's life science community in the research, development, and delivery of innovative products.

Biocom launched its Big Data/Informatics Committee to work collaboratively with stakeholders to solve big data challenges and to maximize opportunities that drive innovation for California. The committee is comprised of senior-level professionals with strong subject matter expertise in big data applications for the life sciences. Some members of the committee have engaged in developments in Europe pertaining to FAIR[1] data.

In our mission of providing feedback and communication between regulators and industry, we are writing in response to the National Institutes of Health's (NIH) request for information (RIF) on the draft Strategic Plan for Data Science.

Biocom applauds the agency's commitment to addressing the key challenges emerging in data science. Data science is an integral component of biomedical research. For industry to continue the development of scalable and robust data science applications, we must address fundamental issues in storing data efficiently and securely; making data reusable to as many people possible; developing a research workforce poised to capitalize on advances in data science and information technology; and setting policies for data use.

[1] Biomedical research data should adhere to FAIR principles, meaning that it should be Findable, Accessible, Interoperable, and Reusable.

Biocom believes that NIH's Strategic Plan for Data Science is strong and reflects a consistent vision with well-defined goals, but NIH could be more explicit on how the strategic plan will be realistically implemented. NIH should also consider including the following concepts to its plan: Internet of FAIR Data and Services; FAIRification of data; machine-readable metadata; FAIR Metadata profiles; and FAIR data stewardship[2]. Lastly, we suggest that NIH use FAIR metrics and FAIR data points as performance measures, which would provide a fine-grained monitoring of the exchange of data and services.

Our comments, detailed in the balance of this letter, will provide our recommendations on how this strategic plan could be implemented.

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

To achieve this goal, NIH states that it will partner with cloud service providers for cloud storage, computational, and related infrastructure services needed to facilitate the deposit, storage, and access to large, high-value NIH data sets. Biocom would like to acknowledge that the global trend is to steer away from massive cloud storage and cloud computing in favor of local storage of FAIR data sets connected through FAIR metadata search engines, as well as bringing the algorithms to the data instead of transferring data.

Local entities can be sponsored to store and maintain similar types of data or data with similar reference points, such as UCSC Genome Browser[3]. The NIH may promote partnerships between cloud service providers and local institutions for reproducibility and at the economies of scale necessary to maintain services in a cost-effective and sustainable manner.

Local entities may also need support to allow academic centers and institutes to steward data transfer from individual laboratories to data repositories. This draws from similar efforts in healthcare, i.e. meaningful use and health information exchanges, where institutes could receive funding for milestones and usage of data-sharing infrastructure to develop the infrastructure and community services necessary for laboratories to share and annotate disparate data. The same effort might also promote data exchange with data repositories to normalize data to the institution's own model.

Additionally, NIH plans to develop strategies to link high-value NIH data systems, building a framework to ensure they can be used together rather than existing as isolated data silos. To link high-value NIH data systems, Biocom believes the agency will need machine-readable metadata and a FAIR Pointer index or registry. It is also recommended that NIH guide the creation of relevant metadata profiles. Furthermore, data standards for metadata tagging across biomedical and clinical research need to be agreed upon and defined so that data access across different database types can be linked.

Goal 2: Promote Modernization of the Data-Resources Ecosystem

To achieve this goal, NIH plans to coordinate and collaborate with other federal, private, and international funding agencies and organizations to promote economies of scales and synergies and prevent unnecessary duplication. The mandate of the Global Open (GO) FAIR International Support and Coordination Office is to prevent duplication and prevent adoption of incompatible standards.

[2] A data steward is a job role that involves planning, implementing and managing the sourcing, use and maintenance of data assets in an organization. Data stewardship is the process for maximizing the value of data as an institutional resource.
[3] The UCSC Genome Browser is an on-line genome browser hosted by the University of California, Santa Cruz (UCSC)

Biocom suggests that NIH collaborate with one or more GO FAIR Support and Coordination Offices in the US to build on these efforts.

NIH plans to refocus its funding priorities on the utility, user service, accessibility, and efficiency of operation of repositories. Biocom recommends requiring data repositories to be FAIR at the source, for example, through the creation of standard machine-readable metadata models for repositories. Biocom also recommends the NIH to fund or support relational schema and crosswalks across metadata to engineer broad utilization of data. These efforts currently are performed by individual laboratories and projects and detract from the resources and time necessary for innovation and research.

We applaud the agency's strategy to distinguish between databases and knowledgebases to improve the evaluation of data repository utility. The distinction between the two is very important because databases contain primary data for analysis while knowledgebases contain books, journals and other published data not appropriate for analysis. Additionally, metrics on the reuse of data is important but information on the use of knowledgebases in which the results of analytics of data are stored is equally relevant and important to determine the 'Return on Investment' on NIH-funded projects.

Biocom is supportive of NIH's strategy to create an environment in which individual laboratories can link datasets to publications in the National Center for Biotechnology Information (NCBI)'s PubMed Central publication databases. Industry would benefit from sharing source data as opposed to high-level summaries, in addition to accessibility to supplementary data supporting conclusions of research published in articles.

To leverage ongoing initiatives like the All of Us Research Program and Cancer Moonshot, NIH plans to integrate patient health data into the biomedical data-science ecosystem in ways that maintain security and confidentiality and are consistent with informed consent, applicable laws, and high standards for ethical conduct of research. Biocom recommends that NIH not only include health data as found in Medicaid/Medicare databases but also data from clinical trials to better understand drug effectiveness.

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

In an effort to optimize the biomedical data-science ecosystem, NIH plans to leverage existing vibrant tool-sharing systems to help establish a more competitive market. Biocom recommends that NIH look to R Bioconductor as an example of openware development for a biomedical data tool because it is a large open-source library of tools used to analyze biological data in a standard format.

NIH also plans to evaluate and fund tool development separately from the support of databases and knowledgebases. The Dutch Ministry of Education and Research, through its GO FAIR International Support and Coordination Office, has an existing roadmap to support this. The elements of the roadmap are development of FAIRification tool kits, FAIR metrics services, FAIR data points, FAIR pointers (i.e. unique identifiers for digital objects), and registries for data model (F100), metadata profiles, ontologies, access protocols, and identifiers. All these elements are crucial in enabling funders to request a FAIR research cycle with mandatory FAIR data stewardship plans and associated FAIR data services payable from an allocated line item in the budget of the grant. The roadmap is currently being implemented through the GO FAIR Implementation Networks and Biocom recommends that the agency review and adopt a similar roadmap to avoid recreating a new system and to maximize its current funding of FAIR-related items, like the FAIR metrics.

NIH aims at promoting the establishment of environments in which high quality, open-source data management, analytics, and visualization tools can be obtained and used directly with data in the NIH Data Commons and/or other cloud environments. Biocom recommends that NIH use the GO FAIR Data Stewardship Wizard and the FAIR metrics as a guideline on high-quality, open-source data management. NIH could use this opportunity to support innovation in deeper search capabilities.

NIH should also work to improve the current framework for reporting data in clinicaltrials.gov. HRB, the Irish Health Research Board funder, and ZonMW, the Dutch funder of Health and Biomedical research, early movers in GO FAIR, are launching pilots that include the creation of machine-readable metadata and FAIR Data Points to track Clinical Research projects and post-project re-use. Biocom also recommends that NIH require mandatory reporting of data to include failed laboratory research data, trials and links to metadata as well as low level data.

Goal 4: Enhance Workforce Development for Biomedical Data Science

NIH plans to develop training to facilitate familiarity and expertise with data-science approaches and effective and secure use of various types of biomedical research data to improve staff knowledge of data science. In addition to knowledge of data science, Biocom believes that NIH should also include in this training curriculum, focus on data FAIRification and FAIR data stewardship. We recommend that NIH collaborate and learn from existing programs like those in GO FAIR's training program, which targets FAIR data stewards and has already been successfully piloted by European funders.

It is essential that we continue to equip the next generation of researchers with the skills needed to foster advancements in human health with data science. Biocom supports NIH's efforts to enhance quantitative and computational training for graduate students and postdoctoral fellows. Biocom suggests that NIH should create grant funding opportunities to expose high school and post-secondary students to data science training and opportunities and encourage them to enter the field.

Goal 5: Enact Appropriate Policies to Promote Stewardship and Sustainability

In Goal 5, NIH states that there is an urgent need to develop clear guidelines for how data must be stored and shared, where and in what form it must be stored, as well as practical solutions for sustaining valuable data resources and determining priorities for data-resource funding. NIH should clarify that guidelines are needed for both humans and machines.

In addition to our comments on the goals, NIH should consider the comments below on how to implement the strategic plan:

- **Recommendation 1**: Align this plan with the NLM Strategic Plan 2017-2027
- **Recommendation 2**: Collaborate with the European Commission on its process for the Implementation Roadmap of the EOSC (European Open Science Cloud) and the GO FAIR kick-start implementation process.
- **Recommendation 3**: Organize a fact-finding workshop between NIH/NLM/GO FAIR to discuss what is already done, available and piloted.
- **Recommendation 4**: Involve industry as early as possible in Public Private Partnerships (PPP). Many companies already provide professional FAIRification services, Data stewardship planning tools, and AI analytics based upon FAIR data.

- **Recommendation 5**: In addition to calls, use other funding instruments to begin the implementation of the plan, such as the European Implementation model of Implementation Networks (PPP collaborations).

This letter focuses on principles on how data should be produced but it is important to acknowledge that further discussion will be warranted to fully address the challenges related to data privacy and security, such as what data should be shared; where it should be share; who should have access; what time frame it should become available and to what extent others can use it. Biocom would like to continue to work with NIH to address these areas.

Thank you again for the opportunity to provide these comments. We look forward to a continued dialogue with the NIH on the Strategic Plan for Data Science. If you have any questions about these comments, please contact Brittany Blocker, Manager of Regulatory Affairs at bblocker@biocom.org.

Sincerely,

Joe Panetta
President and CEO
Biocom