National Institutes of Health

Data and Informatics Working Group

Draft Report to
The Advisory Committee to the Director

June 15, 2012

1 EXECUTIVE SUMMARY

1.1 Committee Charge and Approach

In response to the accelerating growth of biomedical research datasets, the Director of the National Institutes of Health (NIH) charged the Advisory Committee to the Director (ACD) to form a special Data and Informatics Working Group (DIWG). The DIWG was asked to provide the ACD and the NIH Director with expert advice on the management, integration, and analysis of large biomedical research datasets. The DIWG was charged to address the following areas:

- research data spanning basic science through clinical and population research
- administrative data related to grant applications, reviews, and management
- management of information technology (IT) at the NIH

The DIWG met nine times in 2011 and 2012, including two in-person meetings and seven teleconferences, toward the goal of providing a set of consensus recommendations to the ACD at its June 2012 meeting. In addition, the DIWG published a Request for Information (RFI) as part of their deliberations (see Appendix, Section 6.1 for a summary and analysis of the input received).

The overall goals of the DIWG's work are at once simple and compelling:

- to advance basic and translational science by facilitating and enhancing the sharing of research-generated data
- to promote the development of new analytical methods and software for this emerging data
- to increase the workforce in quantitative science toward maximizing the return on the NIH's public investment in biomedical research

The DIWG believes that achieving these goals in an era of "Big Data" requires innovations in technical infrastructure and policy. Thus, its deliberations and recommendations address technology and policy as complementary areas in which NIH initiatives can catalyze research productivity on a national, if not global, scale.

1.2 DIWG Vision Statement

Research in the life sciences has undergone a dramatic transformation in the past two decades. Colossal changes in biomedical research technologies and methods have shifted the bottleneck in scientific productivity from data production to data management, communication, and interpretation. Given the current and emerging needs of the biomedical research community, the NIH has a number of key opportunities to encourage and better support a research ecosystem that leverages data and tools, and to strengthen the workforce of people doing this research. The need for advances in cultivating this ecosystem is particularly evident considering the current and growing deluge of data originating from next-generation sequencing, molecular profiling, imaging, and quantitative phenotyping efforts.

The DIWG recommends that the NIH should invest in technology and tools needed to enable researchers to easily find, access, analyze, and curate research data. NIH funding for methods and equipment to adequately represent, store, analyze, and disseminate data throughout their useful lifespan should be coupled to NIH funding toward generating those original data. The NIH should also increase the capacity of the workforce (both for experts and non-experts in the quantitative disciplines), and employ strategic planning to leverage IT advances for the entire NIH community. The NIH should continue to develop a collaborative network of centers to implement this expanded vision of sharing data and developing and disseminating methods and tools. These

centers will provide a means to make these resources available to the biomedical research community and to the general public, and will provide training on and support of the tools and their proper use.

1.3 Overview of Recommendations

A brief description of the DIWG's recommendations appears below. More detail can be found in Sections 2-4.

Recommendation 1: Promote Data Sharing Through Central and Federated Catalogues

Recommendation 1a. Establish a Minimal Metadata Framework for Data Sharing

The NIH should establish a truly minimal set of relevant data descriptions, or metadata, for biomedically relevant types of data. Doing so will facilitate data sharing among NIH-funded researchers. This resource will allow broad adoption of standards for data dissemination and retrieval. The NIH should convene a workshop of experts from the user community to provide advice on creating a metadata framework.

Recommendation 1b. Create Catalogues and Tools to Facilitate Data Sharing

The NIH should create and maintain a centralized catalogue for data sharing. The catalogue should include data appendices to facilitate searches, be linked to the published literature from NIH-funded research, and include the associated minimal metadata as defined in the metadata framework to be established (described above).

Recommendation 1c. Enhance and Incentivize a Data Sharing Policy for NIH-Funded Data

The NIH should update its 2003 data sharing policy to require additional data accessibility requirements. The NIH should also incentivize data sharing by making available the number of accesses or downloads of datasets shared through the centralized resource to be established (described above). Finally, the NIH should create and provide model data-use agreements to facilitate appropriate data sharing.

Recommendation 2: Support the Development, Implementation, Evaluation, Maintenance, and Dissemination of Informatics Methods and Applications

Recommendation 2a. Fund All Phases of Scientific Software Development via Appropriate Mechanisms

The development and distribution of analytical methods and software tools valuable to the research community occurs through a series of stages: prototyping, engineering/hardening, dissemination, and maintenance/support. The NIH should devote resources to target funding for each of these four stages.

Recommendation 2b. Assess How to Leverage the Lessons Learned from the National Centers for Biomedical Computing

The National Centers for Biomedical Computing (NCBCs) have been an engine of valuable collaboration between researchers conducting experimental and computational science, and each center has typically prompted dozens of additional funded efforts. The NIH should consider the natural evolution of the NCBCs into a more focused activity.

Recommendation 3: Build Capacity by Training the Workforce in the Relevant Quantitative Sciences such as Bioinformatics, Biomathematics, Biostatistics, and Clinical Informatics

Recommendation 3a. Increase Funding for Quantitative Training and Fellowship Awards

NIH-funded training of computational and quantitative experts should grow to help meet the increasing demand for professionals in this field. To determine the appropriate level of funding increase, the NIH should perform a supply-and-demand analysis of the population of computational and quantitative experts, as well as develop a strategy to target and reduce identified gaps. The NCBCs should also continue to play an important educational role toward informing and fulfilling this endeavor.

Recommendation 3b. Enhance Review of Quantitative Training Applications

The NIH should investigate options to enhance the review of specialized quantitative training grants that are typically not reviewed by those with the most relevant experience in this field. Potential approaches include the formation of a dedicated study section for the review of training grants for quantitative science (*e.g.,* bioinformatics, clinical informatics, biostatistics, and statistical genetics).

Recommendation 3c. Create a Required Quantitative Component for All NIH Training and Fellowship Awards

The NIH should include a required computational or quantitative component in all training and fellowship grants. This action would contribute to substantiating a workforce of clinical and biological scientists trained to have some basic proficiency in the understanding and use of quantitative tools in order to fully harness the power of the data they generate. The NIH should draw on the experience and expertise of the Clinical and Translational Science Awards (CTSAs) in developing the curricula for this core competency.

Recommendation 4: Develop an NIH-Wide "On-Campus" IT Strategic Plan

Recommendation 4a. For NIH Administrative Data:

The NIH should update its inventory of existing analytic and reporting tools and make this resource more widely available. The NIH should also enhance the sharing and coordination of resources and tools to benefit all NIH staff as well as the extramural community.

Recommendation 4b. For the NIH Clinical Center:

The NIH Clinical Center (CC) should enhance the coordination of common services that span the Institutes and Centers (ICs), to reduce redundancy and promote efficiency. In addition, the CC should create an informatics laboratory devoted to the development of implementation of new solutions and strategies to address its unique concerns. Finally, the CC should strengthen relationships with other NIH translational activities including the National Center for Advancing Translational Sciences (NCATS) and the CTSA centers.

Recommendation 4c. For the NIH IT and Informatics Environment:

The NIH should employ a strategic planning process for trans-agency IT design that includes considerations of the management of Big Data and strategies to implement models for high-value IT initiatives. The first step in this process should be an NIH-wide IT assessment of current services and capabilities. Next, the NIH should continue to refine and expand IT governance. Finally, the NIH should recruit a Chief Science Information Officer (CSIO) and establish an

external advisory group to serve the needs of/guide the plans and actions of the NIH Chief Information Officer (CIO) and CSIO.

Recommendation 5: Provide a Serious, Substantial, and Sustained Funding Commitment to Enable Recommendations 1-4

The current level of NIH funding for IT-related methodology and training has not kept pace with the ever-accelerating demands and challenges of the Big Data environment. The NIH must provide a serious, substantial, and sustained increase in funding IT efforts in order to enable the implementation of the DIWG's recommendations 1-4. Without a systematic and increased investment to advance computation and informatics support at the trans-NIH level and at every IC, the biomedical research community will not be able to make efficient and productive use of the massive amount of data that are currently being generated with NIH funding.

1.4 Report Overview

This report is organized into the following sections following the executive summary to provide a more in-depth view into the background and the DIWG's recommendations:

Section 2 provides a detailed account of the DIWG's recommendations related to research data spanning basic science through clinical and population research, including workforce considerations (Recommendations 1-3).

Section 3 provides a detailed explanation of the DIWG's recommendations concerning NIH "on campus" data and informatics issues, including those relevant to grants administrative data, NIH CC informatics, and the NIH-wide IT and informatics environment (Recommendation 4).

Section 4 provides details about the DIWG's recommendation regarding the need for a funding commitment (Recommendation 5).

Section 5 includes references cited in the report.

Section 6 includes appendices.