Response to Request for Information:
Public Access to Digital Data Resulting From Federally-Funded Scientific Research

From:
Sage Bionetworks
1100 Fairview Ave. N.
Seattle WA 98109

Summary
It is a critical time to implement and enforce effective rules for sharing federally-sponsored research results. Taxpayers have paid for the work and the entire community should have access.

The United States federal government is the largest funder of medical research in the world. As such its policies have a profound global impact on transforming research into improved healthcare. This is particularly true as we dive into a new world where mega-data genomic technology will require cooperation between data creators, data analysts and medical researchers to support the development of innovative therapies and personalized medicine.

Specific Comments related to the RFI:

(1) *Usable* copies of *all* federally-funded digital data should be placed in a publicly-accessible repository within a short period (30 days) of *creation* as an absolute condition of funding.

(2) Patentable Intellectual property rarely exists in primary digital data. Dissemination of federally funded research data does not normally compromise stakeholder intellectual property interests.

(3) Existing grant review panels have domain experts who can apply discipline-specific criteria.

(4) Long term stewardship is part of the full cost of research and should be handled as part of indirect cost negotiations

(5) Universities can foster the required cultural change intrinsic to rapid sharing of data by requiring and monitoring compliance in the same way that human subjects research is monitored and by using hiring and promotion criteria that include community contributions from data sharing. The former will only occur if required by funding agencies

(6) The real cost of data management should be included in indirect cost agreements.

(7) Applicants should be required to provide public repository ID numbers for every dataset referenced in grant progress reports, renewals and proposals as is currently required for publications.

(8) Funding agencies should support a range of pilot consortia focused on creating community-based platforms for effective storage and use of digital data outside of the traditional investigator-initiated, project-based funding programs.

(9) Data ID systems are a logical starting point for referencing and attributing credit.

(10, 11, 12) Absolute data standards are difficult as technologies evolve rapidly and in unexpected directions. Data repositories must be flexible and promote interoperable formats that allow workflow provenance.

(13) Publications can use data IDs as is already standards in many fields

Medical Research Context: Despite the sequencing of the human genome and tremendous advances in medical technology, genomic research has thus far not made significant contributions to healthcare. A critical factor is the inadequate sharing of genetic data, tools and results required for complex, data-intensive human biology research.

Medical research largely still occurs in isolated labs across the biomedical landscape using data that is not broadly accessible. Results are shared by publication in scientific journals after considerable delay and often lack sufficient detail for reproducibility. Without transparency, the quality of published models cannot be established and the refinement of modeling techniques cannot be pursued. There is also a prevalent attitude that federally-funded scientists who create data somehow own that data. This creates hoarding behaviors that are reinforced by research institutes and universities where credit shared may be promotions lost.

The research community needs to adopt a new set of behavioral norms to fully exploit genomic data where multiple investigators participate in, and are appropriately acknowledged and rewarded for, contributing to common, pre-competitive projects. This evolution needs clear demonstrations that breakthrough research findings require widespread and open community involvement as an incentive to change scientist behaviors. *Revised funding rules will change behaviors*.

Ultimately, progress is dependent on both changes in reward systems and the creation of forums for collaborative modeling with open repositories of curated and readily available and usable genomic data and disease models and tools for data analysis.
The interpretation of genomic data requires innovative computational methodologies. Integrative analyses that combine genotypic and molecular trait data to predict phenotypic outcomes provide a powerful means to improve the mechanistic understanding of disease and to develop novel approaches to treatment. Using computational modeling it is now possible to provide frameworks that describe complex physiological systems and predict the causal relationships between molecular states and disease. Early applications of these approaches have provided a boost in the understanding of disease pathologies. Despite the tremendous potential of this field, there are significant barriers to success related to the public availability of “usable” large-scale genomic data, the rapid development, evaluation and deployment of good methods to analyze the data, and the availability of disease models to drive downstream experimental validation and therapeutic development.

Integrative genomic analysis requires a move towards an open source, community-based modeling environment. It is a big challenge and no one organization has the resources to succeed in isolation. Rapid sharing of digital data and results is a prerequisite to make effective use of the large investments being made in genomic research. The success of many open-source software projects has demonstrated that distributed, decentralized and appropriately incentivized teams can effectively collaborate on complex and large-scale projects in an appropriate framework. The software industry has many examples of how such approaches has been transformative. Many of the most widely-used software projects are available for free and are open source (e.g., Android OS, Apache server, Firefox). Such infrastructure is increasingly hosted and managed by third party providers (e.g., SourceForge, GitHub, Google Code). This has allowed development teams to focus on truly novel areas thereby spawning entire new businesses (Facebook, Twitter, etc.). It has also transformed

the way in which software engineers receive recognition for their work and advancement in their career. For example a GitHub profile is beginning to complement or even replace a traditional resume in the software industry as it links directly to an individual's contributions on hosted projects.

Genomic science and public healthcare need ready access to digital data as well as open forums to compare, refine and deploy new methods for the analysis of high dimensional population-level genomic data.

Submission prepared by;
Jonathan Izant PhD
Vice President, Sage Bionetworks

About Sage Bionetworks: Sage Bionetworks is a 501(c)(3) nonprofit biomedical research organization created to change how researchers approach the complexity of human biological information and the treatment of disease.

Sage Bionetworks' mission has five interdependent themes:

- Research on computational network models of disease
- Pilot projects trialing disruptive models of research cooperation
- Rules and rewards that promote data sharing and collective research
- Building the computational platform for a digital Commons
- Activating public engagement and access

We are driving a cultural change around the elimination of disease by activating patients, shifting scientists to share the data and models needed to build better models of disease. To do this, we are building an open Commons called 'Synapse' where data can be shared and a compute space where predictive disease models can be co-evolved so that industry and academia can jointly benefit from understanding biology.

http://www.sagebase.org info@sagebase.org