

Brian M. Bot



Sage Bionetworks - 1100 Fairview Ave N • Seattle, WA 98109
E-Mail: brian.bot@sagebase.org Web: www.sagebase.org

Date: January 12, 2012

To: Office of Science and Technology (OSTP), on behalf of
National Science and Technology Council (NSTC)
digitaldata@ostp.gov

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

Open access to biomedical research is essential. It will provide an increased return on the federal government's investment and ultimately will result in improved patient care.

From the onset of my career as a biostatistician, I was sheltered from the struggles plaguing biomedical research. Fortunate to have my first job in the innovative and intellectually stimulating atmosphere of the Mayo Clinic, my perspective on the public access conversation was limited to the tiresome albeit valid argument as to whom "owns" data resulting from federally funded research. This debate merely scratches the surface of the issue of public access to research data.

Current infrastructure is not equipped to handle the data deluge that is omnipresent in the research setting. Even mighty research institutions are being forced to rethink the way data is managed. Startups, innovators, and changemakers don't stand a chance.

Now biomedical research has reached its tipping point.

The research community has benefited from open access to digital data with the Human Genome Project (HGP) providing a forward thinking model of public access to valuable research data. The guiding principles set out in the Bermuda Accord undoubtedly avoided much duplication of efforts and thus years of wasted time, effort, and resources.

Other examples such as the Gene Expression Omnibus (GEO), the Database of Genotypes and Phenotypes (dbGaP), and The Cancer Genome Atlas (TCGA) provide value as well, but are flawed by usability and accessibility issues.

Arguably, half-baked public data resources are worse than not making the data available at all. Usable data in standardized formats should be required, not just the act of making data available. Governance inhibiting access of "public" research data presents unnecessary obstacles for researchers trying to reuse data already made available.

Data alone is never enough. Metadata describing an experiment from beginning to end is as important as the raw data itself. Data without links to clinical characteristics make it impossible to build predictive models of disease. Data without experimental confounders can bias model building and confuse downstream interpretation. I would like to see a push for not only public access to digital data, but also public access to full experimental information. The distinction is subtle, but I feel necessary for the true value of open access to be realized.



The logistics of making research data, especially genomic data, widely consumable is no small task. Instituting policies that provide infrastructure, or incentivize the building of infrastructure, can help ease the burden. By building a genomic information superhighway, great opportunity become available for a burgeoning sector of the economy dealing with digital data stewardship.

It is evident that policies seen as optional or mandatory (but loosely enforced) will not be prioritized. Motives of the Individual do not align with the greater good. Individual researchers have many more reasons to conceal their research data than to make it broadly available and useful.

Sadly, scientists are a stubborn bunch. The scientific culture is stale, and simple policy shifts could help to transform it in the blink of an eye. The very scientists living on the bleeding edge of innovation in their respective niche fields ironically lag in the technological advances that allow other sectors to flourish. Transparency, crowdsourcing, and meta-analytics have proven useful but have not been adopted by the masses. Open access policies will force researchers down these advantageous paths.

It is my hope that the spirit of the HGP and Bermuda Accord will continue to live on through policy statements within the National Institute of Health (NIH), but an appropriate incentive structure to encourage those ideological policies is severely lacking. Current incentive structures for career scientists, especially in academia, do not reward altruistic endeavors. There is no place on a CV for having provided readily accessible, usable, and valuable data for the community at large. Measures such as publication rate, grant funding levels, and ultimately tenure are no longer good surrogates for a researcher's impact.

Great opportunities are being missed. The longer public access to digital data is delayed, the less opportunity our creative free market system is able to take advantage of the wealth of taxpayer-funded resources that are currently locked up. Lowering the barriers of entry to usable research data is a proven catalyst for innovation and entrepreneurship. If we can further encourage this type of environment through policy changes, our next generation of researchers and the scientific community as a whole can reap the benefits. Imagine the collaborative atmosphere and exponential return on investment for federal research funding that would be possible if we succeed.

It is my hope that publicly informed and expert driven policy choices can shape this promising revolution. The opportunity is present to create a synergistic movement that points all researchers, public and private, towards common goals that will benefit the greater good and ultimately more rapid progress toward important scientific advances.

Regards,

Brian M. Bot