



# AMERICAN UNIVERSITY

W A S H I N G T O N , D C

Jorge L. Contreras  
202-274-4124  
contreras@wcl.american.edu

January 12, 2012

National Science and Technology Council (NSTC)  
Office of Science and Technology Policy (OSTP)  
Attention: Ted Wackler, Deputy Chief of Staff  
Via email: [digitaldata@ostp.gov](mailto:digitaldata@ostp.gov)

Re: OSTP Request for Information: Public Access to Digital Data Resulting from Federally Funded Scientific Research (76 Fed. Reg. No. 214 at 68,517 (Nov. 4, 2011))

Dear Mr. Wackler:

I appreciate the opportunity to share comments with OSTP regarding public access to digital data resulting from federally-funded scientific research. I am a professor of law at American University, prior to which I spent seventeen years as a practicing attorney representing major research institutions, R&D consortia and private enterprises engaged in technical and scientific work. I have recently served as a member of the National Advisory Council on Human Genome Research and currently serve as Co-Chair of the National Conference of Lawyers and Scientists and Co-Chair of the American Bar Association Section of Science & Technology Law's Committee on Technical Standardization. My current research focuses on the production and dissemination of scientific and technical information.

Responses to specific items of the OSTP RFI are set forth below and represent my own views, and not those of American University, Washington College of Law, or any of the other organizations mentioned above.

### *Preservation, Discoverability, and Access*

**(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?**

Public access to data from federally-funded research has several potential benefits. These include enabling scientists to reproduce and validate the results of their peers, reducing the incidence scientific fraud and misrepresentation, and accelerating the overall progress of scientific discovery. These benefits, however, do not come without cost, and several counterbalancing effects of data release must be considered. These include the exposure of potentially

identifiable personal information of human research subjects, the reduction of intellectual property protection in released data, and the reduction of publication opportunities for data-generating scientists.

In previous work, I have analyzed these costs and benefits in the context of federal data release policies relating to human genomics research, and have traced the development of these policies from 1992 through 2009.<sup>1</sup> Over this period, genomics data release policies have evolved significantly. The so-called “Bermuda Principles”, adopted in 1996, required the release of genomic sequence data within 24 hours of generation. Among the purposes of the Bermuda Principles was to limit the ability, both of data producers and third parties, to obtain patents claiming raw sequence data generated by the public genome project. Despite the success of the initial genome project, subsequent projects and policies have seen a measured retreat from the sweeping requirements of Bermuda. Today, the major federal genomics data release policies (e.g., the 1997 NIH GWAS policy, the 2008 ENCODE and modENCODE policies, and the 2009 Human Microbiome Project policy) require that data be released (i.e., deposited into publicly-accessible, federally-managed databases such as Genbank and dbGaP) subject to detailed user agreements and requirements (see also response to Question 7, below). These agreements typically impose an “embargo” period on users of data, prohibiting them from publishing or presenting conclusions derived from this data during a pre-determined period of time (usually 9-12 months).

Interestingly, a number of private biomedical research projects have adopted similar approaches to data release. In some of these cases, however, data is withheld for a period of time (also between 9-12 months), after which it is released without restriction. The fact that both public and private research projects have independently arrived at “embargo” periods in roughly the same range (i.e., 9-12 months) implies that the length of this period (which I refer to as the “latency” period) can be viewed as an equilibrium of sorts. That is, at the latency equilibrium point, the various stakeholder groups negotiating such policies (i.e., funders, data-generating scientists, data-using scientists, and public advocates) are each willing, albeit reluctantly at times, to make data public. A shorter period would not be acceptable to data-generating scientists as it would not adequately compensate them for their effort, and a longer period would not be acceptable to funders and data-using scientists, who have an interest in making such data freely available at the earliest possible time. Thus, as a result of multilateral compromise, an equilibrium latency period acceptable to all stakeholder groups may emerge.

We have seen the development of similar latency periods and equilibria in the area of scientific publishing. In this area, publishers, funders, libraries, universities and scientists have engaged in a series of explicit and implicit negotiations (contractual, administrative and legislative) over the appropriate latency period before which published scientific research may be released to the public free from access restrictions. Again, a latency period between 6-12 months emerges as an

---

<sup>1</sup> See Jorge L. Contreras, *Prepublication Data Release, Latency, and Genome Commons*, 329 SCIENCE 393 (2010), *Data Sharing, Latency Variables and Science Commons*, 25 BERKELEY TECH. L.J. 1601 (2010), Jorge L. Contreras, *Bermuda’s Legacy: Patents, Policy and the Design of the Genome Commons*, 12 MINN. J.L. SCI. & TECH. 61 (2011).

equilibrium point in several independent contexts. And again, this convergence suggests that a latency period in this range appropriately rewards publishers, on one hand, and adequately meets the access requirements of scientists, libraries and the public, on the other hand.

In my view, these data suggest that “market” forces (i.e., the negotiation and interplay of interested stakeholder groups) can arrive at protective periods that are substantially shorter than default intellectual property rules (20 years in the case of patents, and nearly 100 years in the case of copyright).<sup>2</sup>

Extrapolating beyond the areas of genomics and scientific publishing, I believe that many (if not all) fields of scientific endeavor will exhibit a latency equilibrium point at which the release of data to the public will duly balance the costs and benefits to the relevant stakeholder groups. I do not suggest that this equilibrium point will be the same in all fields. In fact, I believe that different fields could exhibit radically different latency equilibria. It is quite likely that a field such as paleoanthropology would exhibit a substantially longer latency equilibrium point than genomics. Moreover, I do not attempt here to pre-judge whether the release of data prior to or after publication of the associated analysis is preferable in one field versus another,<sup>3</sup> and again suspect that the norms, practices and practicalities of different scientific disciplines would dictate the optimal practice in each such discipline.

Nevertheless, I believe that attempting to discern the latency equilibrium point in a scientific field can yield valuable information regarding the appropriate balancing of intellectual property and other interests among competing stakeholder groups, and I would encourage OSTP to consider this analysis in its future policy development activity.

**(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?**

Improving compliance with Federal data access policies embodies two distinct and significant challenges: monitoring compliance and enforcing (and formulating) penalties for noncompliance.

*Monitoring:* To date, most violations of data access policies are identified on an ad hoc basis by scientific colleagues, competitors and journal editors (whistleblowers) rather than agency staff. Given Federal budgetary constraints, I imagine it unlikely that an effective inter-agency

---

<sup>2</sup> The fact that these observed latency equilibrium periods are so much shorter than default intellectual property periods also suggests that the default periods may, in some cases, be unnecessarily lengthy (a point that has been made by many others).

<sup>3</sup> Pre-publication and post-publication data release, and the considerations surrounding each, are analyzed in a pair of companion pieces that appeared in *Nature* in 2009. Toronto International Data Release Workshop Authors, *Prepublication Data Sharing*, 461 NATURE 168 (2009) and Paul N. Schofield, Tania Bubela, Thomas Weaver, et al., *Post-Publication Sharing of Data and Tools*, 461 NATURE 171 (2009).

compliance monitoring system could be implemented in the near future. A reasonable alternative might be to coordinate (and incentivize) private whistleblowing activity through a central Federal data oversight board. Such a board could promulgate rules and guidance regarding the reporting of data access/use violations and could encourage relevant stakeholders to report such violations on an anonymous basis. To the extent that investigation of reports is warranted, the board could allocate resources to such investigation or refer the reported violation to the relevant agency.

*Enforcement:* I have previously written about the somewhat dubious enforceability of Federal data access policies, particularly with respect to third party users outside of the U.S. There are several potential avenues toward improving policy enforceability, both contractually and through international trade mechanisms. However, before embarking on a major enforceability program, it would be prudent to gather data regarding compliance as outlined above. In particular, information regarding overall levels of non-compliance, together with any data that emerge regarding the characteristics of policy violators and the nature, frequency and seriousness of their violations, would be useful to consider. With this information in hand, one could more accurately assess options for improving compliance and potential penalties for non-compliance.

### *Standards for Interoperability, Re-Use and Re-Purposing*

**(10) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?**

Data exchange standards are ubiquitous in today's technology-driven marketplace. Such standards, which are commonplace in the information, computing and telecommunications (ICT) sector, specify the manner in which products and services offered by different vendors interact with one another. A number of these standards, including WiFi, USB, CD, DVD, PDF and HTML, have become household terms, and thousands of others ensure that a vast array of products and services connect and communicate seamlessly in a manner that is largely invisible to the consumer.

Standardization in the ICT sector, however, has not come without a cost. Over the past two decades, the industry has been plagued by lawsuits brought by participants in the standards-development process as well as by government regulators and affected third parties. Two types of claims generally arise in standards-related litigation: claims that the standards process has been abused to disadvantage one or more companies ('process-abuse' claims) and claims that a participant in the standards-development process has improperly asserted patents against an implementer of the standard ('patent hold-up' claims). Standards-development organizations in the ICT sector have responded to these claims by promulgating rules and policies of increasing sophistication, both to specify procedures designed to avoid abusive activity and to accommodate the requirements of participants who control significant patent assets.

The explosion of data-driven scientific research over the past two decades has led to a surge of interest in the development of interoperability and compatibility standards. These range from standards for genome annotation and controlled vocabularies (ontologies) to data formats and search engine integration. A variety of organizations are involved in these standards-development activities, from large, established standards bodies such as the Institute for Electrical and Electronics Engineers (IEEE) and the Worldwide Web Consortium (W3C) to broad-based industry associations such as the European Bioinformatics Institute (EBI) to narrowly-focused efforts such as the Proteomics Standards Initiative (PSI) and the Functional Genomics Investigation Ontology (FuGO) project. To date, most science-driven standardization efforts have been free of the litigation that has plagued the ICT industry. But with the increasing adoption of standards by researchers and vendors, the issues faced by ICT standards groups will become increasingly relevant.

Today, the large majority of science-focused standards-development efforts are relatively informal and unstructured, and are thus ill-equipped to address or deter process abuse and patent hold-up scenarios. In many cases, the organizations responsible for standards development either lack written policies entirely, or adopt vague, aspirational statements regarding a desire that materials produced be “open” and publicly-available. This informal and minimalist approach not only invites opportunistic behavior, but also leaves aggrieved participants with little legal recourse after abuse has occurred.<sup>4</sup>

Accordingly, I recommend that OSTP encourage science-focused standards-development organizations to review their existing policies and procedures with care. To the extent that they fail to address key points regarding process openness and intellectual property, these policies and procedures should be supplemented.<sup>5</sup> For example, if it is the desire of a group that all scientists worldwide be permitted to access and implement a new scientific data sharing standard without the payment of copyright or patent licensing fees, the group’s policy should state this clearly and require contributing participants to commit not to assert copyrights or patents in connection with the standard. Hopefully, such modest prophylactic measures will enable the scientific standards community to avoid the disruptive and costly litigation that has affected the ICT sector.

---

<sup>4</sup> See Jorge L. Contreras, *Legal Issues in the Development of Biological Research Standards*, 26 NATURE BIOTECHNOLOGY 498 (2008).

<sup>5</sup> A number of resources exist to assist non-lawyers with understanding and developing appropriate standards-development policies. See ABA COMM. ON TECH. STANDARDIZATION, STANDARDS DEVELOPMENT PATENT POLICY MANUAL (Jorge L. Contreras, ed., 2007).

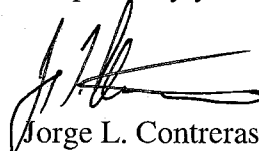
Contreras Response to OSTP RFI on Scientific Data

January 12, 2012

Page 6

Thank you again for the opportunity to offer these comments in response to your inquiries. Please do not hesitate to let me know if there is any additional information that I can provide in support of these matters.

Respectfully yours,

A handwritten signature in black ink, appearing to read 'J. Contreras', with a long horizontal flourish extending to the right.

Jorge L. Contreras