SJ Quinney College of Law, University of Utah
Utah Law Digital Commons

Utah Law Faculty Scholarship | Utah Law Scholarship

12-2020

U.S. Federal Genomic Data Release and Access Policies

Jorge L. Contreras

Follow this and additional works at: https://dc.law.utah.edu/scholarship

 Part of the Intellectual Property Law Commons

Forthcoming in *Bioinformatics, Medical Informatics and the Law* (Jorge L. Contreras, A. James Cuticchia & Gregory Kirsch, eds., Edward Elgar: 2021)

U.S. FEDERAL GENOMIC DATA RELEASE AND ACCESS POLICIES

By: Jorge L. Contreras

Introduction

Researchers today have access to a vast aggregation of human and nonhuman genomic data, largely on an open access basis. According to the Joint Genome Institute's Genomes OnLine Database (GOLD), data from more than 40,000 sequencing projects around the world, representing more than 375,000 different organisms, were publicly available to researchers as of July 2020.[1] The availability of this tremendous public resource is due, in large part, to the data release policies developed a quarter century ago, toward the beginning of the Human Genome Project (HGP), which have been carried forward, in modified form, to the present. These policies impose requirements on both the generators of data (typically the sequencing centers and other laboratories conducting genetic experiments) and the users of that data (i.e., researchers who download and/or use it). This article, briefly outlines the history of such data release policies, particularly in the U.S. and with respect to human genomic data, and provides an overview of the obligations imposed on both data generators and data users.

The Genomic Data Landscape

The principal global databases for the deposit of genomic sequence data are GenBank, which is administered by the National Center for Biotechnology Information (NCBI) a division of the NIH's National Library of Medicine (NLM), the European Molecular Biology Library (EMBL) in Hinxton, England, and the DNA Data Bank of Japan (DDBJ). NCBI also maintains the RefSeq database, which consolidates and annotates much of the sequence data found in GenBank. In addition to DNA sequence data, genomic studies generate data relating to the association between particular genetic markers and disease risk and other physiological traits. This type of data, which is more complex to record, search and correlate than raw sequence data, is stored in databases such as NLM's Database of Genotypes and Phenotypes (dbGaP).

Bermuda and the Origins of Rapid Genomic Data Release (1996)

In the 1980s, U.S. science funding agencies began to require that researchers working on federally-funded projects release the data generated by their research to the public 6-12 months after collection, or at the latest when associated results were published in the literature (generally 12-24 months after data collection).[2] Along these lines, when the HGP was first organized, NIH

[1] For an overview of different genomic data resources comprising the "genomic commons", see Jorge L. Contreras & Bartha M. Knoppers, *The Genomic Commons*, 19 ANNUAL REV. GENOMICS & HUMAN GENETICS 429 (2018).

[2] *See, e.g.,* National Academies of Science, Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age 64 (2009); National Research Council, Sharing Publication-Related Data And Materials: Responsibilities Of Authorship In The Life Sciences 75 (2003); National Research Council, Bits Of Power – Issues In Global Access To Scientific Data 80–82 (1997).

Electronic copy available at: https://ssrn.com/abstract=3692635

and DOE developed formal guidelines for the sharing of HGP data.[3] These guidelines required that data generated by the HGP be deposited in GenBank, making it available to all scientists worldwide within six months after generation.

In 1996 the HGP's sequencing of the human genome was scheduled to begin. That year, approximately fifty HGP project leaders from around the world met in Hamilton, Bermuda to deliberate over the speed with which HGP data should be released to the public, and whether the 6-month "holding period" approved in 1992 should continue.[4] The resulting "Bermuda Principles" established that all DNA sequence information from large-scale human genomic sequencing projects should be "freely available and in the public domain in order to encourage research and development and to maximize its benefit to society."[5] Specifically, the Bermuda Principles require that all human genomic sequence assemblies greater than one kilobase (Kb) in length be released *within twenty-four hours* after assembly, and that finished annotated sequences should be submitted *immediately* to a public database.

The Bermuda Principles were revolutionary in that they established, for the first time, that data from public genomic projects should be released to the public almost immediately after their generation. While this requirement seemed radical at first, it has since become ingrained as part of NIH's basic position treating genomic data as a public good that should be widely available and unencumbered.[6]

Genomic Data Release Post-HGP: The Ft. Lauderdale Principles (2003)

Initial drafts of the human genome sequence were published by the HGP and Celera Genomics, a competing private effort, in 2001, and the full human genome sequence was finalized by the HGP in 2003.[7] That year, the Wellcome Trust convened a meeting in Ft. Lauderdale, Florida to revisit rapid data release issues in the "post-genome" world.[8] While the Ft. Lauderdale participants "enthusiastically reaffirmed" the 1996 Bermuda Principles, they also expressed concern over the inability of data generating scientists to study their results and publish analyses prior to the public release of data. The most significant outcome of the Ft. Lauderdale meeting was a consensus that the Bermuda Principles should apply to each "community resource project" (CRP), meaning "a research project specifically devised and implemented to create a set of data, reagents or other material whose primary utility will be as a resource for the broad scientific

[3] *NIH, DOE Guidelines Encourage Sharing of Data, Resources*, HUMAN GENOME NEWS (Oak Ridge Nat'l Laboratory, Oak Ridge, Ten.), Jan. 1993, at 4.

[4] *International Large-Scale Sequencing Meeting*, HUMAN GENOME NEWS (Oak Ridge Nat'l Laboratory, Oak Ridge, Ten.), Apr.-June 1996, at 19.

[5] *Summary of Principles Agreed at the First International Strategy Meeting on Human Genome Sequencing*, U.S. DEPARTMENT OF ENERGY GENOME PROGRAM, http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml.

[6] For a comprehensive history of the Bermuda Principles and their impact on genomic science, see Katherine Maxson Jones, Rachel A. Ankeny, Robert Cook-Deegan, *The Bermuda Triangle: The Pragmatics, Policies, and Principles for Data Sharing in the History of the Human Genome Project*, 51 J. HIST. BIOLOGY 693 (2018). For the Bermuda Principles' impact on patenting human genetic material, *see* Jorge L. Contreras, *Bermuda's Legacy: Patents, Policy and the Design of the Genome Commons*, 12 MINNESOTA J. L., SCI. AND TECH. 61 (2011).

[7] *See, generally*, Francis Collins, *Opinion: Has the Revolution Arrived?*, 464 NATURE 674 (2010).

[8] Report of Meeting organized by the Wellcome Trust, *Sharing Data from Large-Scale Biological Research Projects: A System of Tripartite Responsibility* (Jan. 14–15, 2003), *available at* http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf.

Electronic copy available at: https://ssrn.com/abstract=3692635

community." Under this definition, the twenty-four hour rapid release rules of Bermuda would be applicable to large-scale projects generating non-human sequence data, other basic genomic data maps, and other collections of complex biological data such as protein structures and gene expression information. In order to effectuate this data release requirement, funding agencies were urged to designate appropriate efforts as CRPs and to require, as a condition of funding, that rapid, pre-publication data release be required in such projects.

Notwithstanding this show of support, the Ft. Lauderdale participants acknowledged that rapid, pre-publication data release might not be feasible or desirable in all situations, particularly for projects other than CRPs. In particular, the notion of a CRP, the primary goal of which is to generate a particular data set for general scientific use, is often distinguished from "hypothesis-driven" research, in which the investigators' primary goal is to solve a particular scientific question, such as the function of a specific gene or the cause of a specific disease or condition.[9] In hypothesis-driven research, success is often measured by the degree to which a scientific question is answered rather than the completion of a quantifiable data set. Thus, the early release of data generated by such projects would generally be resisted by the data generating scientists who carefully selected their experiments to test as yet unpublished theories. Giving such data away before their theories are published could potentially enable a competing group to "scoop" the originating group, a persistent fear among highly competitive scientists.

The NIH GWAS Policy and dbGaP (2007)

Following the completion of the HGP, researchers began to conduct large-scale studies seeking to associate particular sets of genetic markers with disease risk and other physiological traits (genome-wide association studies or GWAS). GWAS data, which necessarily includes phenotypic and clinical data in addition to genomic data, is more complex to record, search and correlate than raw sequence data, is stored in databases such as NLM's Database of Genotypes and Phenotypes (dbGaP). In response to the growing number of GWAS being conducted and the large amount of data generated by such studies, in August 2007 the NIH released a new policy regarding the generation, protection and sharing of data generated by all federally-funded GWA studies.[10]

The NIH GWAS Policy requires that researchers submit data collected from GWA studies to dbGaP, maintained by the NLM. In addition to genomic data, dbGaP can also accommodate phenotypic data, which includes elements such as de-identified subject age, ethnicity, weight, demographics, drug exposure, disease state, and behavioral factors, as well as study documentation and statistical results, including linkage and association analyses. Given the potential sensitivity of phenotypic data, dbGaP allows access to data on two levels: open and controlled. Open data access is available to the general public via the Internet and includes non-sensitive summary data, generally in aggregated form. Data from the controlled portion of the database may be accessed

[9] Jane Kaye, et al., *Data Sharing in Genomics – Re-shaping Scientific Practice*, 10 NATURE REV. GENETICS 331, 332 box 1 (2009).

[10] Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS), 72 Fed. Reg. 49290, 49294–97 (Aug. 28, 2007) (hereinafter "NIH GWAS Policy"); Modifications to Genome-Wide Association Studies (GWAS) Data Access, NAT'L INST. OF HEALTH (Aug. 28, 2008) (hereinafter "GWAS Amendment").

Electronic copy available at: https://ssrn.com/abstract=3692635

only under conditions specified by the data supplier, often requiring certification of the user's identity and research purpose.

The 2007 policy requires that researchers deposit descriptive information about each GWA study for inclusion in the "open access" portion of dbGaP. Researchers are also "strongly encouraged" to submit study results, including phenotypic, exposure and genotypic data, for inclusion in the "controlled access" portion of the database "as soon as quality control procedures have been completed."[11]

One of the principal concerns with GWAS data was the risk that phenotypic or clinical information could eventually be traced back to individuals.[12] To address this concern, the NIH GWAS Policy requires that GWAS data be de-identified in accordance with HIPAA guidelines.[13] Moreover, the data in the controlled-access portion of dbGaP may be released only after approval of the proposed research use by a Data Access Committee (DAC),[14] and then only under a signed Data Use Certification. The Data Use Certification requires researchers and their institutions to agree, among other things, to use data only for the approved research purpose, to protect data confidentiality, to implement appropriate data security measures, not attempt to identify individual data subjects, not to sell any data, not to share data with third parties, and to report violations to the DAC. Finally, the GWAS Policy states NIH's position that a request under the Federal Freedom of Information Act (FOIA)[15] for the release of individually-identifiable GWAS data would constitute an "invasion of personal privacy", and will be denied.[16]

The NIH GWAS Policy was amended in August, 2008, following the publication of a scientific paper demonstrating that inferences regarding individual identity could be made using statistical techniques.[17] Due to concerns relating to potential identification of GWAS subjects, NIH withdrew certain GWAS-generated SNP data from the publicly-accessible portions of dbGaP and certain NCI databases and placed them in the controlled-access portions of these databases.[18]

The NIH GWAS Policy addresses the publication priority concerns of data generators by stating an expectation that users of GWAS data refrain from submitting their analyses and conclusions for publication, or otherwise presenting them publicly, during an "exclusivity" period of up to twelve months from the date that the data set is made available. NIH also expresses a "hope" and expectation that "genotype-phenotype associations identified through NIH-supported and NIH-maintained GWAS datasets and their obvious implications will remain available to all investigators, unencumbered by intellectual property claims."[19] It goes on to explain that "[t]he filing of patent applications and/or the enforcement of resultant patents in a manner that might restrict use of NIH-supported genotype-phenotype data could diminish the potential public benefit they could provide."[20] However, in an effort to show some support for patent seekers,

---

[11] NIH GWAS Policy, *supra* note 10, at 49295.
[12] *Id*. at 49292.
[13] *Id*. at 49295.
[14] *Id.* at 49296.
[15] Federal Freedom of Information Act, 5 U.S.C. § 552 (2006).
[16] FOIA Exemption 6, 5 U.S.C. §552(b)(6).
[17] GWAS Amendment, *supra* note 10; Nils Homer et al., *Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays*, PLoS GENETICS (Aug. 2008).
[18] GWAS Amendment, *supra* note 10.
[19] NIH GWAS Policy, *supra* note 10, at 49296.
[20] *Id*. at 49297

Electronic copy available at: https://ssrn.com/abstract=3692635

the GWAS Policy also "encourages patenting of technology suitable for subsequent private investment that may lead to the development of products that address public needs."[21]

Project-Specific Genomic Data Release Policies

In the years following the Ft. Lauderdale meeting, numerous large-scale genomic research projects were launched with increasingly sophisticated requirements regarding data release. These policies implement their requirements through contractual mechanisms that are more tailored and comprehensive than the broad policy statements of the HGP era. Moreover, improvements in database technology enabled the provision of differentiated levels of data access, the screening of user applications for data access, and improved tracking of data access and users.

In addition, the proliferation of genomic databases and research projects funded by NIH expanded the role of NIH's NCBI in genomic research. Rather than acting as a passive funder of genomic research, NIH and its sub-institutes have taken an increasingly active role in the development, curation and dissemination of data created through NIH-funded research.[22]

Genetic Association Information Network (GAIN) (2006)

The Genetic Association Information Network (GAIN) was established in 2006 by the Foundation for the National Institutes of Health (FNIH), the NIH and several corporations.[23] GAIN's purpose was to conduct GWA studies of the genetic basis for six common diseases. Data generators in the GAIN program were required to sign an applicant agreement agreeing to various program commitments, including "immediate" release of data generated by the project.[24] Over the course of the three-year project, approximately 18,000 human DNA samples were genotyped.[25] The resulting data was deposited in dbGaP. Researchers wishing to access data from the controlled portion of the database were required to register with, and be approved by, the GAIN Data Access Committee (DAC).[26]

Perhaps most importantly, the GAIN policy was the first U.S. Federal genomic data release policy to introduce a temporal restriction on the *users* of the data (as opposed to the temporal release requirements imposed on data *generators* by the Bermuda Principles). That is, in order to secure a period of exclusive use and publication priority for the data generators, data users were prohibited from submitting abstracts and publications and making presentations based on GAIN data for a specified embargo period.[27] The duration of the embargo period for a given data set is

---

[21] *Id.* at 49296.
[22] *See* Jorge L. Contreras, *Leviathan in the Commons: Biomedical Data and the State*, in GOVERNING MEDICAL KNOWLEDGE COMMONS, Ch. 2 (Katherine Strandburg, Brett Frischmann, Michael Madison eds., Cambridge Univ. Press: 2017).
[23] *See generally* The GAIN Collaborative Research Group, *New models of collaboration in genome-wide association studies: the Genetic Association Information Network*, 39 NATURE GENETICS 1045 (2007) [hereinafter GAIN Consortium].
[24] GAIN Consortium, *supra* note 23, at 1048 (Box 1).
[25] Teri A. Manolio, *Collaborative genome-wide association studies of diverse diseases: programs of the NHGRI's office of population genomics*, 10 PHARMACOGENOMICS 235, 236 (2009).
[26] GAIN Consortium, *supra* note 23, at 1049.
[27] *Id*.

Electronic copy available at: https://ssrn.com/abstract=3692635

identified in the relevant data repository and may vary by data set, but has generally been set at nine months.[28]

The Cancer Genome Atlas (TCGA) (2006)

In 2006, the National Cancer Institute (NCI) and NHGRI launched a project to catalog genomic changes relating to cancer.[29] The Cancer Genome Atlas (TCGA) project generated genomic sequence and related data for cancer tumor cells, but also kept track of a large amount of clinical data, including patient diagnosis, treatment history and ongoing status. Due to the specialized nature of the project data, deposits were made both in dbGaP as well as a TCGA-specific database administered by NCI. Given the potential for identifying individual patients from their genomic and phenotypic data, attention was paid to controlling access to TCGA data. TCGA data is available in an open-access tier and a controlled-access tier similarly to dbGaP.[30] Open-access is provided for data that cannot be aggregated to generate an individually-identifiable dataset, whereas controlled-access enables researchers to access clinical and individually-unique data. Access to the controlled-access data tier requires the user's acknowledgement of a Data Access Certification containing restrictions on research use, security, transferability and other matters.

ENCODE and modENCODE (2007)

In 2007 NIH launched the ENCODE and modENCODE projects to identify functional genomic elements in humans and two model organisms.[31] The ENCODE data release policy[32] designates the project as a "community resource project", but also recommends a nine-month embargo period during which users of released data are requested not to publish or present results based on that data. The ENCODE Policy distinguishes between published and unpublished data, verified and unverified data, and offers several examples of the data use implications for different types of studies. The length and complexity of the policy evidences the agency's and the participants' desire for clear guidelines and the avoidance of misunderstandings regarding the release of data, as the diversity of participants, organisms and data types had expanded dramatically beyond those originally considered by the framers of the Bermuda Principles.

The Human Microbiome Project (2008)

The Human Microbiome Project (HMP) was a large-scale community resource project initiated in 2008 that was designed to identify and sequence the genomes of selected microorganisms inhabiting the human body.[33] While much HMP data was subject to rapid

[28] *Id.*
[29] Francis S. Collins & Anna D. Barker, *Mapping the Cancer Genome*, SCI. AM., Mar. 2007, at 50.
[30] *Data Access*, THE CANCER GENOME ATLAS DATA PORTAL, http://cancergenome.nih.gov/dataportal/data/access/.
[31] Susan E. Celniker et al., *Unlocking the Secrets of the Genome*, 459 NATURE 927 (2009).
[32] ENCODE Consortia, DATA RELEASE, DATA USE, AND PUBLICATION POLICIES (2008), *available at* http://www.genome.gov/Pages/Research/ENCODE/ENCODEDataReleasePolicyFinal2008.pdf.
[33] Peter J. Turnbaugh, et al., *The Human Microbiome Project*, 449 NATURE 804 (2007); *Human Microbiome Project Awards Funds for Technology Development, Data Analysis and Ethical Research*, NIH NEWS (Oct. 7, 2008), http://www.genome.gov/27528386.

Electronic copy available at: https://ssrn.com/abstract=3692635

Bermuda-like disclosure requirements, investigators were permitted to withhold certain other data from the public for a period of several months.[34] This hold-back period was intended to permit HMP researchers to analyze and prepare publications on their data before it is released to competing researchers. The reasons that researchers, who are often driven by intense competitive pressure to publish and claim credit for discoveries, pushed for such hold-back periods is clear. However, it also appears, at least in the case of HMP, that NIH did not vigorously advance the patent deterrent arguments that previously motivated policy decisions during the HGP and its immediate aftermath.

Data Release Policies Beyond Genomics

The success and broad adoption of genomics data release policies incorporating the Bermuda and Ft. Lauderdale Principles have led scientists in related fields to consider the adoption of analogous principles in their own research. One prominent example occurred in 2008, when the National Cancer Institute convened a meeting of proteomics researchers in Amsterdam to "identify and address potential roadblocks to rapid and open access to [proteomics] data."[35] Participants identified technical, infrastructure and policy challenges to the rapid release of proteomic data. Technical challenges included the wide variety of disparate platforms and techniques used to generate proteomic data, making "raw" data from experimental instruments difficult to interpret by scientists unfamiliar with, or lacking access to, the instruments used to generate the data. Proteomics also lacks the established public database infrastructure of genomics. Whereas DNA sequence data can be deposited readily in GenBank, the EMBL or DDBJ, and is often deposited in all three, there is no common public data repository for proteomic data, and existing proteomic databases suffer from inconsistent and sometimes incompatible data formats. Finally, unlike genomics, in which the entire field focused for several years on the single HGP project, proteomics research lacks a unifying policy core and proteomics-focused journals have each developed their own, sometimes inconsistent, guidelines for data submission.

Notwithstanding these difficulties, the Amsterdam participants articulated six data release and sharing principles that reflect the spirit of the Bermuda and Ft. Lauderdale Principles: (1) Timing (should depend on the nature of the effort generating the data, but should in no event be later than publication or, for community resource projects, following appropriate quality assurance procedures), (2) Comprehensiveness (full raw data sets should be released together with associated metadata and quality data), (3) Format (standardized formats are encouraged), (4) Deposition to repositories (central repositories for proteomic data should be established), (5) Quality metrics (central repositories should develop metrics for assessing data quality), and (6) Responsibility (scientists, funding agencies and journals share responsibility for ensuring adherence to community data release standards).

In 2009, more than one hundred scientists, journal editors, legal scholars and representatives of governmental and private funding agencies met in Toronto to assess the current state of rapid pre-publication data release and the applicability of the Bermuda Principles in

[34] *HMP Data Release and Resource Sharing Guidelines for Human Microbiome Project Data Production Grants*, NIH COMMON FUND, http://commonfund.nih.gov/hmp/datareleaseguidelines.asp.
[35] Henry Rodriguez et al., *Recommendations From the 2008 International Summit on Proteomics Data Release and Sharing Policy: A Summit Report*, 8 J. PROTEOMICS RES. 3689 (2009).

Electronic copy available at: https://ssrn.com/abstract=3692635

projects well beyond the generation of genomic sequence data.[36] The participants reaffirmed a general community commitment to rapid pre-publication data release, expanding the scope of projects as to which these principles should apply to all biomedical datasets having "broad utility, are large in scale … and are 'reference' in character". Specifically, they cited, in addition to genomic and proteomic studies, structural chemistry, metabolomics and RNAi datasets as well as annotated clinical resources such as cohorts, tissue banks and case-control studies.

The expansion of rapid pre-publication data release principles beyond genomics and proteomics projects, which often have as their ultimate goal the generation of a large data set, to these other areas necessarily raises issues concerning the appropriateness of rapid data release in hypothesis-driven research. Accordingly, the Toronto participants concurred that, while funding agencies should *require* rapid pre-publication data release for "broad utility" projects, rapid data release "should not be mandated" for projects that are generally hypothesis-driven. The Toronto participants also addressed the priority concerns of data generators versus data users, observing anecdotally that in many cases data users have, in fact, published papers based on publicly-released data sets *before* the publication of the data generators' papers analyzing the data sets themselves, and that this situation caused no "serious damage" to the data generators' subsequent publications. Nevertheless, the participants acknowledged the acceptability of a "protected period" during which data users could be restricted from publishing on released data sets, cautioning, however, that this period should never exceed one year. The Toronto participants produced a set of "best practices" embodying these principles and applying them to the three constituencies originally identified in Ft. Lauderdale – funding agencies, data generators and data users – as well as to scientific journals, which were urged to monitor and provide guidance relating to data release issues.

Obama Administration Directives On Federal Data Sharing

In February 2013, John Holdren, Director of the White House Office of Science and Technology Policy (OSTP) issued a memorandum directing Federal agencies with annual R&D budgets of more than $100 million to develop plans for increasing public access to the results of research that they fund.[37] The Memorandum recognized that "making research results accessible to the largest possible audience – other researchers, business innovators, entrepreneurs, teachers, students, and the general public – can boost the returns from Federal investments in R&D. Increased access expands opportunities for new scientific knowledge to be applied to areas as diverse as health, energy, environmental protection, agriculture, and national security and to catalyze innovative breakthroughs that drive economic growth and prosperity."

In May 2013, President Obama issued Executive Order 13642, "Making Open and Machine Readable the New Default for Government Information,"[38] which required that data generated or funded by Federal agencies be made available in open, machine-readable formats while appropriately safeguarding privacy, confidentiality, and security. Shortly thereafter, the White House Office of Management and Budget (OMB) issued a new Federal Open Data Policy (M-13-

[36] Toronto International Data Release Workshop Authors, *Prepublication Data Sharing,* 461 Nature 168 (2009).
[37] https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
[38] White House Executive Order 13642, "Making Open and Machine Readable the New Default for Government Information,
https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Electronic copy available at: https://ssrn.com/abstract=3692635

13)[39] that designated data as a valuable national resource and strategic asset. The Federal Open Data Policy requires that Federal agencies make data and information open in machine-readable format in order to "accelerate scientific breakthroughs and innovation, promote entrepreneurship, and enhance economic growth and job creation."

NIH's Fourth Generation Data Sharing Policies

The Genomic Data Sharing (GDS) Policy (2014)

In 2014, responding to the OMB's 2013 directive, NIH released a new Genomic Data Sharing (GDS) policy. The GDS policy expanded the scope of the 2007 GWAS policy in several important respects. Whereas the GWAS policy covered only human GWAS data, the GDS Policy covers all large-scale human and non-human genomic data. This comprehensive scope unifies NIH's approach to different data types across its many institutes. The GDS policy also requires, among other things, that human subjects provide informed consent to the broad sharing and research use of data generated using their DNA, a requirement that is likely to result in significant debate and to require further clarification from NIH as it is implemented.

Under the GDS policy, human genomic data must be submitted to NIH promptly following cleaning and quality control (generally within three months after generation). Once submitted, this data may be retained by NIH for up to six months prior to public release. Non-human and model organism data, on the other hand, can be retained by data producers until their initial analyses of the data are published, representing a much longer lead time. In both cases, once released, data is not subject to further embargoes or restrictions on analysis or publication. The GDS policy thus diverges from the GWAS and other contemporary policies in that it (a) permits the withholding of data from the public for a fixed period of time, and (b) does not utilize embargoes on data usage following its release.[40]

The Cancer Moonshot Public Access and Data Sharing (PADS) Policy (2017)

In 2016, the National Cancer Institute (NCI), the largest research funding organization within NIH, launched an ambitious $1.8 billion research program directed at accelerating cancer research – the Cancer Moonshot Program. In 2017, the Cancer Moonshot Program adopted a new Public Access and Data Sharing (PADS) Policy to encourage the broad sharing of data and publications among researchers and the public.[41]

The Policy addresses both open access to publications resulting from Cancer Moonshot-funded research as well as the release of data generated by that research.[42] With regard to data, the

[39] Exec. Off. Pres., Off. Mgt. & Budget, Open Data Policy-Managing Information as an Asset, Memorandum M-13-13, May 9 2013, https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2013/m-13-13.pdf

[40] For a discussion and critique of these changes, see Jorge L. Contreras, *NIH's Genomic Data Sharing Policy: Timing and Tradeoffs*, 31 TRENDS IN GENETICS 55 (2015).

[41] National Cancer Institute (NCI) (2017a). NCI Cancer Moonshot Public Access and Data Sharing Policy (Aug. 4, 2017).

[42] Release of scientific publications is beyond the scope of this paper. For a discussion of NIH's open access publication requirements, *see, e.g.*, Jorge L. Contreras, *Confronting the Crisis in Scientific Publishing: Latency, Licensing and Access*, 53 SANTA CLARA L. REV. 491 (2013); Tammy M. Frisby & Jorge L. Contreras, *The NCI Cancer*

Electronic copy available at: https://ssrn.com/abstract=3692635

PADS Policy states that "to the extent feasible," data should be made publicly available simultaneously with the publication of research results. Unlike prior NIH data sharing policies that pertained primarily to genomic and related data, the PADS Policy defines data that must be shared as any "recorded factual material commonly accepted in the scientific community as necessary to document and support research findings in Publications". This broad definition thus encompasses all forms of clinical, pharmacological, demographic, analytical, survey, and other data that might be collected or developed as part of a Cancer Moonshot project.

Because the scope of covered data is so broad, NCI offers applicants for Cancer Moonshot funding a degree of flexibility in deciding how to release their research results and data. Thus, the only firm requirement of the PADS policy is that applicants submit to NCI a written plan that describes a proposed process for making data "immediately and broadly available to the public" and, if such sharing is not possible, a justification for why it is not. NCI emphasizes that its review of grant funding proposals "will give funding priority to those Applicants that submit an appropriate [PADS] [p]lan that ensures maximal sharing of [publications and data] arising from the award." Despite its ambitious goals, one recent study has found that few submitted PADS plans during the first year that the PADS policy was in effect conformed to its requirements.[43]

NIH's Draft Policy for Data Management and Sharing (DMS) (2019)

In 2015, NIH began planning a more sweeping overhaul of its trans-institute data sharing program (NIH, 2015). In November, 2019, it released a draft Policy for Data Management and Sharing (DMS) for public comment.[44] Like the Cancer Moonshot PADS Policy, the draft DMS policy mandates that researchers seeking NIH funding submit a data sharing plan as part of their grant applications. This plan would be reviewed by the funding agency in evaluating the application, and compliance would be monitored during the course of the funded research program. A final DMS policy has not been released as of this writing.

OPEN Government Data Act (2019)

In January 2019, the U.S. Congress enacted the Open, Public, Electronic, and Necessary (OPEN) Government Data Act,[45] codifying the open data principles contained in the OSTP, OMB and Executive memoranda of 2013. Among other things, the OPEN Government Data Act requires Federal agencies to publish data arising from Federal research online using standardized, machine-readable data formats, with metadata uploaded to the Data.gov catalog, and to prepare inventories of their available datasets.

Moonshot Public Access and Data Sharing (PADS) Policy – Initial Assessment and Implications, 2 DATA & POLICY E9 (2020).

[43] *See* Frisby & Contreras, *supra* note 42.

[44] National Institutes of Health (NIH). Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance. 84 Fed. Reg. 60398 (Nov. 8, 2019).

[45] P.L. 115-435, 132 Stat. 5529 (Jan. 14, 2019) (Title II of the Foundations for Evidence-Based Policymaking Act of 2018).

Electronic copy available at: https://ssrn.com/abstract=3692635

Conclusion

The forward-looking genomic data release policies developed by the HGP a quarter century ago have shaped the landscape of scientific data release both in the U.S. and around the world.[46] Today, data sharing from government-funded scientific projects is not limited to the field of genomics, but it is a Federal statutory requirement for data arising from all scientific fields of inquiry from astronomy to geology to biochemistry to psychology. Yet despite the endorsement of data sharing principles at the highest levels of the United States government, it is not clear that researchers in all scientific fields are prepared to embrace rapid and public data sharing as Federal policy requires. The priority and publication concerns of data-generating researchers are still very real, particularly in fields outside of genomics,[47] and as a recent study of the latest NCI data sharing policy shows, there may be a lack of understanding of data sharing practices and principles even in related biomedical research fields.[48] Thus, despite Federal policy on the books, additional training, education and socialization of data sharing norms within the broader scientific community must continue in order to realize the full potential of this key element of open science.

[46] *See, generally*, Contreras & Knoppers, *supra* note 1, Yann Joly et al., *Open science and community norms: Data retention and publication moratoria policies in genomics projects*, 12 MEDICAL L. INTL. 92 (2014).
[47] *See* Rudolf I. Amann, et al., *Toward unrestricted use of public genomic data*, 363 SCIENCE 350 (2019).
[48] *See* Frisby & Contreras, *supra* note 42.

Electronic copy available at: https://ssrn.com/abstract=3692635