INFORMATION ACCESS

Prepublication Data Release, Latency, and Genome Commons

Scientific information commons can be modulated by adjusting the twin "dials" of knowledge and rights latency.

Jorge L. Contreras

Researchers must disclose their data in order to achieve recognition and to enable others to test, validate, and challenge their hypotheses. In doing so, they create bodies of shared knowledge that are analogous to traditional public resources, such as forests and freeways, often referred to as "commons" (*1*, *2*). The rate at which data are added to these information commons, however, varies greatly. The traditional practice has been to contribute experimental and observational data to the commons when, or shortly after, the analysis of that data is published, sometimes years after its initial collection (*3*, *4*). Because of busy schedules, competitive pressures, and other interpersonal vagaries, the sharing of scientific data can be inconsistent even after publication (*5*, *6*). Many traditional data-sharing practices were challenged, with significant and lasting effect, during the race to sequence the human genome.

The Bermuda Principles

In 1992, shortly after the initiation of the international Human Genome Project (HGP), the U.S. National Institutes of Health (NIH) and Department of Energy (DOE) adopted a policy requiring that genomic data be released within 6 months after generation (*7*), a substantial reduction from the typical 12- to 18-month period required by other government-funded projects (*3*, *8*). But by 1996, leaders of the HGP realized that even 6 months was too slow for the massive project. They convened in Bermuda and agreed that all HGP sequence data be released to public databases (e.g., NIH's GenBank), without restrictions on subsequent publication, within 24 hours after generation (*9*).

At a practical level, the so-called "Bermuda Principles" were necessary to coordinate activities of more than a thousand HGP researchers around the world. But other motivations were also at work. To many researchers, including some HGP leaders, the human genome sequence was fundamentally different from other large data sets: It represented the common heritage of the human species and should not be encumbered by patents, a growing concern in the mid-1990s (*10*–*12*). By requiring rapid, prepublication data release of the human genome sequence, the ability of both the publicly funded data generators and others to patent the resulting data would be limited (*13*–*15*). And, whereas delays built into previous policies gave data generators time to analyze results and prepare papers before the data became available to others, the Bermuda organizers sacrificed investigators' publication "priority" in the service of these other program goals.

Latencies: Knowledge and Rights

The framework for the information commons established by the Bermuda Principles and subsequent policies in the genome sciences can be thought of in terms of built-in temporal "latencies." "Knowledge latency," the period between the generation of a particular datum and its mandated release to the public commons (*16*), is immediately followed by "rights latency," the period between the entry of a datum into the commons and the lifting of substantial encumbrances on its use, such as those imposed by policy or contract.

Latency characteristics of U.S. genomic data release policies (*17*) have evolved since Bermuda (see the figure below). A number of factors contribute to the changes in these policies. For example, (i) concerns about academic recognition and tenure "credit" have led to a greater demand for attribution among data generators and the desire for at least some priority in preparing publications based on "their own" data (*18*, *19*). (ii) Investigators whose work is driven primarily by the generation and testing of hypotheses have proven to be less comfortable with rapid, prepublication data release than investigators whose work has been driven primarily by the generation of large data sets (*18*–*20*). And (iii), issues regarding patenting of genetic material have taken on increasing prominence, together with the recognition by U.S. funding agencies that their ability to impose outright prohibitions on patenting by grant recipients is limited by the Bayh-Dole Act of 1980 (*21*, *22*).

To address these issues, genomics policy-makers, after experimenting with a variety of less-specific directives (*23*), began to adapt latency mechanisms to dictate how and when data should be released and used (*25*). Distinct approaches soon emerged, employing latency-based rules to promote policy goals (e.g., access to data to promote scientific advancement, publication priority to provide incentives and to reward researchers, and minimization of encumbrances on the use of data once entered into the commons). But the approaches prioritized these goals differently.

A path taken by NIH is exemplified by the 2007 policy adopted for all NIH-funded genome-wide association studies (GWAS) (*26*) (see the figure). Rapid release of federally funded GWAS data is required, but data users must refrain from publishing or presenting related results during an "embargo" period of up to 12 months. This embargo, which may be implemented via "click-through" agreements or funding policy obligations (*27*), gives data generators a "head start" on preparing publications based on their data, yet data are still broadly available for the general advancement of science. Because the Bayh-Dole Act limits the government's ability

LATENCY VARIABLES IN GENOMIC DATA RELEASE POLICIES

Genomics agreement	Year	Knowledge latency	Rights latency	References
Funded by the U.S. federal government				
NIH-DOE Guidelines	1992	6 months	0	(*6*)
Bermuda Principles	1996	24 hours	0	(*8*)
International HapMap Project	2003	Short	0	(*38*,*39*)
GAIN	2006	Short	9 months	(*40*)
NIH GWAS Policy	2007	Short	12 months	(*26*)
Encode + modENCODE	2008	Short	9 months	(*41*)
Nongovernmental				
SNP Consortium	1998	3 months	0	(*42*)
International SAE Consortium	2007	12 months	9 months	(*28*)

Latency variables in genomic data release policies. Policies designated "short" do not specify an exact timeline for data release but do require that data be released as rapidly as possible after generation or validation.

School of Law, Washington University in St. Louis, St. Louis, MO 63130, USA. E-mail: jlcontreras@wulaw.wustl.edu

Published by AAAS

to impose restrictions directly on patenting, this rapid data release requirement presents a means to frustrate patents that might otherwise restrict use of the data.

An alternate approach has been taken by privately organized groups such as the International Serious Adverse Events Consortium (iSAEC). The consortium establishes a holding period of up to 12 months before data are made publicly accessible, during which data generators can exclusively analyze and prepare papers (*28*). Compared with the NIH GWAS policy, this approach offers a stronger preservation of data generator publication priority, but at the expense of overall scientific advancement (as the data are not available for public use during the 12-month priority period). Patent issues are addressed via a separate "defensive patenting" policy [a strategy not available to governmental agencies (*29*)], under which the consortium files patent applications on discoveries, thus establishing an early "priority date," but will subsequently contribute these rights to the public.

Beyond Genomics

The latency-based framework derived from genome commons policies has broader applicability to other scientific disciplines. Scientific information commons are under discussion in areas ranging from microbiology (*30*) to global climate change (*31*) to molecular chemistry (*32*). Policy designers in these fields cannot adopt wholesale the approaches taken in the genome commons. For example, whereas it may have been appropriate for genome commons policies to limit patent encumbrances on human DNA sequence data, such an approach may be inappropriate for other data sets, such as proprietary chemical compounds. Thus, the unique attributes, history, and norms of a given discipline must shape its data-sharing paradigms.

Issues associated with science commons have also been highlighted by the debate over open-access publishing and the potential release of vast quantities of published and unpublished scientific information to the public. As with genome commons, policy-makers negotiating the contours of open-access databases have already begun to rely on timing mechanisms to balance the competing policy objectives of scientists, government, and publishers (*33*, *34*). The key to developing successful information commons is striking the appropriate balance among these competing objectives. Commons weighted too heavily in favor of data users are not likely to attract sufficient contributions from data generators, whereas commons weighted too heavily in favor of data generators may not optimally advance the interests of science or the public. Thoughtful setting of temporal latency variables will result in scientific information commons that attract optimal quantities of data and thereby serve the overall advancement of science.

References and Notes:

1. C. Hess, E. Ostrom, *Understanding Knowledge as a Commons: From Theory to Practice* (MIT Press, Cambridge, 2007).
2. M. J. Madison, B. M. Frischmann, K. J. Strandberg, *Cornell Law Rev.* **95**, 657 (2010).
3. National Research Council, *Bits of Power: Issues in Global Access to Scientific Data* (National Academy Press, Washington, DC, 1997).
4. The optimal structure of publication-based data release is itself a large and contentious issue that overlaps with, but is beyond the scope of, this paper on prepublication data release. See, e.g., (*35*).
5. D. Blumenthal *et al.*, *Acad. Med.* **81**, 137 (2006).
6. E. G. Campbell *et al.*, *JAMA* **287**, 473 (2002).
7. DOE, *Human Genome News*, January 1993, p. 4.
8. National Academy of Sciences, *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age* (National Academy Press, Washington, DC, 2009).
9. Summary of principles agreed at the International Strategy Meeting on Human Genome Sequencing, Bermuda, 25 to 28 February 1996; www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml#1.
10. National Research Council, *Mapping and Sequencing the Human Genome* (National Academy Press, Washington, DC, 1988).
11. F. Collins, *The Language of Life* (HarperCollins, New York, 2010).
12. J. D. Watson, *Science* **248**, 44 (1990).
13. International Human Genome Sequencing Consortium, *Nature* **409**, 860 (2001).
14. National Research Council, *Reaping the Benefits of Genomic and Proteomic Research* (National Academies Press, Washington, DC, 2006).
15. Releasing data early prevents the data generator from patenting it (there is a statutory ban on patenting material that one has previously released to the public) and also prevents others from patenting it (as the early release acts as "prior art," limiting subsequent claims of "novelty").
16. Most "publicly accessible" genomic and phenotypic data residing in databases such as NIH's Database of Genotypes and Phenotypes (dbGaP) are actually made available only to verified scientific researchers (whether commercial or academic) and not to the public at large. Even under the most stringent rapid data release policy designs, practical, technological, and personnel issues may bear on the usability and effectiveness of data available in the commons.
17. Detailed data release policies exist for non-U.S. genomics projects (e.g., the Wellcome Trust and Medical Research Council in the U.K. and Genome Canada). This article focuses on U.S. policies because they have made great use of the timing mechanisms that are the subject of this analysis. The reason for this may be the lesser prevalence and enforcement of bioscience patents outside the United States, as the laws of those jurisdictions are less amenable to patent filings on genomic discoveries.
18. Toronto International Data Release Workshop Authors, *Nature* **461**, 168 (2009).
19. J. Kaye, C. Heeney, N. Hawkins, J. de Vries, P. Boddington, *Nat. Rev. Genet.* **10**, 331 (2009).
20. National Human Genome Research Institute (NHGRI), NIH, "Policy for release and database deposition of sequence data" (NHGRI, Bethesda, MD, 21 December 2000); www.genome.gov/10000910.
21. A. K. Rai, R. S. Eisenberg, *Law Contemp. Probl.* **66**, 289 (2003).
22. ENCODE Project Data Release Policy (2003–2007), www.genome.gov/12513440.
23. For example, the accord reached at a 2003 summit in Ft. Lauderdale provides for no formal restrictions on the use of data but requests that users "act responsibly to promote the highest standards of respect for the scientific contribution of others" (*24*).
24. Wellcome Trust, Sharing Data from Large-Scale Biological Research Projects: A System of Tripartite Responsibility, meeting, Ft. Lauderdale, FL, 14 to 15 January 2003; www.genome.gov/Pages/Research/WellcomeReport0303.pdf.
25. The temporal mechanisms described in this paper are not the only means that have evolved to address the issue of data generator priority. Other techniques include explicit agency guidelines requiring the acknowledgment of relevant data generators and the early publication by data generators of so-called "statements of intent" and "marker papers."
26. NIH, Policy for sharing of data obtained in NIH supported or conducted genome-wide association studies (GWAS). *Fed. Regist.* **72**, 49,290 (2007).
27. Both the enforceability and actual enforcement of these contractual mechanisms are uncertain. Absent a strong trend toward underenforcement, which has not been observed, it is useful to analyze these policy timing mechanisms at face value, without accounting for underenforcement by rights holders.
28. International SAE Consortium (iSAEC), Intellectual Property Program, www.saeconsortium.org/?q=node/12.
29. Because iSAEC is not directly regulated by the Bayh-Dole Act that limits the government's ability to restrict patents, it has the liberty to modify the patenting process directly, independently of latency mechanisms. Such consortia can thus use timing mechanisms in a more focused manner, without regard for patent concerns, to better address other priorities (e.g., they do not have to rely on the suboptimal embargo approach (*27*) to protect data generators' publication priority).
30. C. Hess, E. Ostrom, *Int. Soc. Sci. J.* **58**, 335 (2006).
31. O. Heffernan, *Nature* Online, 4 September 2009.
32. G. Brumfiel, *Nature* **453**, 139 (2008).
33. NIH, "Revised policy on enhancing public access to archived publications resulting from NIH-funded research" (NOT-OD-08-033, NIH, Bethesda, MD, 7 April 2008).
34. The NIH open-access policy (*33*) gives scientific journals a 1-year "head start" before journal articles are released to public governmental databases. Legislation proposed in the U.S. Congress (*36*) would expand the NIH's open-access mandate to all federal agencies, with a shortened embargo period of 6 months. Legislation opposing both initiatives has also been proposed in Congress (*37*).
35. P. N. Schofield *et al.*, *Nature* **461**, 171 (2009).
36. Federal Research Public Access Act (FRPAA), H.R. 5037 (2010).
37. Fair Copyright in Research Works Act, H.R. 801 (2010).
38. International HapMap Consortium, *Nature* **426**, 789 (2003).
39. International HapMap Project, Data Release Policy, www.hapmap.org/datareleasepolicy.html.
40. Genetic Association Information Network (GAIN), Data Use Certification Agreement (version of 3 December 2008); http://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?view_pdf&stacc=phs000016.v1.p1.
41. NHGRI, "ENCODE consortia data release, data use, and publication policies" (NHGRI, Bethesda, MD, 2008); www.genome.gov/Pages/Research/ENCODE/ENCODEDataReleasePolicyFinal2008.pdf.
42. A. Holden, *Biotechniques* **32**, S22 (2002).

43. J.L.C. is a member of the National Advisory Council for Human Genome Research, the oversight body for NHGRI, NIH; is a paid consultant to the International SAE Consortium Ltd.; and has served as a paid consultant to the SNP Consortium Ltd.

10.1126/science.1189253