November 2006

Harnessing and Sharing the Benefits of State-Sponsored Research: Intellectual Property Rights and Data Sharing in California's Stem Cell Initiative

Arti K. Rai
Duke University Law School

Rebecca Sue Eisenberg
University of Michigan Law School

HARNESSING AND SHARING THE BENEFITS OF STATE-SPONSORED RESEARCH: INTELLECTUAL PROPERTY RIGHTS AND DATA SHARING IN CALIFORNIA'S STEM CELL INITIATIVE

By Rebecca S. Eisenberg & Arti K. Rai†

TABLE OF CONTENTS

I. INTRODUCTION

The considerable attention that the California Institute for Regenerative Medicine (CIRM) and its Independent Citizens' Oversight Committee (ICOC) have already devoted to framing their intellectual property (IP)

---

† Rebecca Eisenberg, Robert & Barbara Luciano Professor of Law, University of Michigan Law School. Arti K. Rai, Professor of Law, Duke Law School. The authors thank Krisnahu Saha for his comments and gratefully acknowledge the support of the National Human Genome Research Institute and the Department of Energy under Grant No. 5P50 G003391-02.

policies[1] is a sure sign of the growing salience of IP in biomedical research. In its Intellectual Property Policy for Non-Profit Organizations (IPPNPO), CIRM has endorsed a "core principle" to "encourage broad dissemination of CIRM-funded intellectual property of all types beyond practices commonly used in 2005 to promote scientific progress."[2] At the same time, CIRM has acknowledged competing interests that might limit such sharing, such as bringing scientific advances to the public through commercialization and providing a financial benefit to the State of California through revenue sharing.[3] Indeed, the text of Proposition 71, the initiative that created CIRM, explicitly sets forth these conflicting interests.[4]

When it comes to balancing interests, the devil is in the details. The IPPNPO is richly detailed with respect to patenting, licensing, and the exchange of research materials. For these matters, the policy generally follows evolving standards of "best practices" for federally-funded research, as articulated in reports from the National Institutes of Health (NIH).[5] For data sharing, however, while it states CIRM's general expectations, the IPPNPO barely touches upon the details.[6]

In recent years, data sharing has been a recurring focus of struggle within the biomedical research community as improvements in informa-

---

1. *See* CIRM, INTELLECTUAL PROPERTY POLICY FOR NON-PROFIT ORGANIZATIONS (2006), *available at* http://www.cirm.ca.gov/policies/pdf/ippnpo.pdf [hereinafter IPPNPO]; *see also* CALIFORNIA COUNCIL ON SCIENCE & TECHNOLOGY, POLICY FRAMEWORK FOR INTELLECTUAL PROPERTY DERIVED FROM STEM CELL RESEARCH IN CALIFORNIA: INTERIM REPORT TO THE CALIFORNIA LEGISLATURE, GOVERNOR OF THE STATE OF CALIFORNIA, CALIFORNIA INSTITUTE FOR REGENERATIVE MEDICINE (2005), *available at* http://www.ccst.us/ccst/pubs/ip/ip%20interim.pdf.

2. IPPNPO, *supra* note 1, at 25.

3. *Id.* at 4-5.

4. California Secretary of State, *Text of Proposed Laws – Proposition 71*, *in* CALIFORNIA OFFICIAL VOTER INFORMATION GUIDE 147 (2004), *available at* http://www.cirm.ca.gov/prop71/pdf/prop71.pdf [hereinafter *Proposition 71*].

5. *See, e.g.*, Principles and Guidelines for Recipients of NIH Grants and Contracts on Obtaining and Disseminating Biomedical Research Resources, 64 Fed. Reg. 72,090 (Dec. 23, 1999), *available at* http://ott.od.nih.gov/pdfs/64FR72090.pdf (cited with approval in IPPNPO, *supra* note 1, at 12).

6. The IPPNPO embraces the lofty aspirations for data sharing set forth in a series of recent reports from the National Research Council. *See, e.g.*, NAT'L RESEARCH COUNCIL, REAPING THE BENEFITS OF GENOMIC AND PROTEOMIC RESEARCH: INTELLECTUAL PROPERTY RIGHTS, INNOVATION, AND PUBLIC HEALTH (2006), *available at* http://newton.nap.edu/catalog/11487.html#toc; NAT'L RESEARCH COUNCIL, SHARING PUBLICATION-RELATED DATA AND MATERIALS: RESPONSIBILITIES OF AUTHORSHIP IN THE LIFE SCIENCES (2003), *available at* http://newton.nap.edu/catalog/10613.html#toc (cited in IPPNPO, *supra* note 1, at 26-27) [hereinafter SHARING DATA & MATERIALS].

tion technology and digital networks have expanded the ways in which data can be produced, disseminated, and used.[7] Electronic archives aggregate data from multiple sources, making it simpler and easier to share data.[8] Such sharing and aggregation facilitate observations that would otherwise be impossible, but data disclosure poses a dilemma for scientists. Data have long been scientists' stock in trade, lending credibility to their claims while highlighting new questions that merit future research funding. Some disclosure is necessary in order to claim these benefits, but data disclosure may also benefit one's research competitors. Scientists who share their data promptly and freely may find themselves at a competitive disadvantage relative to free riders in the race to make future observations and thereby earn further recognition and funding. The possibility of commercial gain further raises the competitive stakes. As information technology has advanced, and as commercial interests in biomedical research have grown, this dilemma has become more pronounced.

The role of statutory IP law in data sharing has been limited. Data per se are generally considered ineligible for either copyright or patent protection.[9] As a consequence, the Bayh-Dole Act,[10] which gives recipients of federal funding broad discretion to seek patent rights in the results of their federally-sponsored research, does not directly address the dissemination of unpatentable data.[11] Meanwhile, the scientific community has sought to clarify its data sharing norms and to determine how to implement them.[12]

---

7. *See, e.g.*, SHARING DATA & MATERIALS, *supra* note 6. *See generally* NAT'L RESEARCH COUNCIL, BITS OF POWER: ISSUES IN GLOBAL ACCESS TO SCIENTIFIC DATA (1997), *available at* http://newton.nap.edu/catalog/5504.html#toc.

8. Of course, integration of data from sources that use different formats can be a problem. But software tools, such as BioPerl in the case of the genomic data produced by the Human Genome Project, can help to address the problem. *See* Colin Crossman & Arti Rai, A Brief History of BioPerl (working paper, on file with authors).

9. For a review of the limits on copyright protection of data with citations to the relevant cases and literature, see J.H. Reichman & Paul F. Uhlir, *A Contractually Reconstructed Research Commons for Scientific Data in a Highly Protectionist Intellectual Property Environment*, 66 LAW & CONTEMP. PROBS. 315, 336-41 (2003). For a review of the limits on patent protection of data, see U.S. Patent & Trademark Office, *Interim Guidelines for Examination of Patent Applications for Patent Subject Matter Eligibility*, 1300 OFFICIAL GAZETTE OF THE U.S. PATENT & TRADEMARK OFFICE 4 (2005), *available at* http://www.uspto.gov/go/og/2005/week47/patgupa.htm.

10. Act of Dec. 12, 1980, Pub. L. No. 96-517, 94 Stat. 3015 (codified as amended at 35 U.S.C. §§ 200-212 (1994)).

11. Although *sui generis* database protection has been enacted in Europe, Council Directive 96/9 of 11 March 1996 on the Legal Protection of Databases, 1996 O.J. (L 77) 20, and proposed in the U.S., it has not yet been passed into law in the U.S. For a review of U.S. database protection proposals from the perspective of the scientific community, see J.H. Reichman & Paul F. Uhlir, *Database Protection at the Crossroads: Recent De-*

One important focus of debate has been the extent of data disclosure that should accompany scientific publication.[13] Although disclosure of research results is the essence of publication, scientific print journals typically reveal data only in summary form. This format provides authors substantial control over access to the underlying raw data. In an earlier era, such summary disclosures may have been necessary as a practical matter, given scarcities of space in print media. Now, however, with the growth of computer networks and information technology, a researcher can easily make vast data sets available over the internet at minimal cost. Yet, a recent survey found that less than half of the most frequently cited journals in the life sciences and medicine had policies requiring deposit of data associated with published articles.[14]

Debate within the scientific community over the disclosure obligations of publishing scientists reached a fevered pitch with the publication of an article in the prestigious *Science* magazine announcing the completion of the human genome sequence by scientists at the private firm Celera.[15] Although Celera made its sequence data available free of charge from its own website, access was restricted along certain dimensions, including quantitative limitations on the amount of data that could be downloaded, a prohibition on redistribution, and additional limitations on commercial users.[16]

The National Research Council of the elite National Academy of Sciences[17] entered into the debate by forming a Committee on Responsibilities of Authorship in the Biological Sciences to examine the topic of sharing published data and materials. The Committee issued a report that called upon authors to include in their publications or otherwise make freely available "the data, algorithms, or other information that is central or integral to the publication—that is, whatever is necessary to support the

---

velopments and Their Impact on Science and Technology, 14 BERKELEY TECH. L.J. 793 (1999).

12. *See, e.g.*, NAT'L RESEARCH COUNCIL, FINDING THE PATH: ISSUES OF ACCESS TO RESEARCH RESOURCES (1999), *available at* http://newton.nap.edu/catalog/9629.html#toc [hereinafter FINDING THE PATH]; SHARING DATA & MATERIALS, *supra* note 6.

13. *See, e.g.*, FINDING THE PATH, *supra* note 12; SHARING DATA & MATERIALS, *supra* note 6.

14. SHARING DATA & MATERIALS, *supra* note 6, at 33 tbl.2-1.

15. J. Craig Venter et al., *The Sequence of the Human Genome*, 291 SCI. 1304 (2001).

16. Science Online, Accessing the Celera Human Genome Sequence Data, http://www.sciencemag.org/feature/data/announcement/gsp.dtl (last visited July 6, 2006).

17. Membership in the National Academy of Sciences is restricted to those scientists who have made highly significant contributions in their fields.

major claims of the paper and would enable one skilled in the art to verify or replicate the claims."[18] The report further indicated that authors should provide data "in a form on which other scientists can build with subsequent research."[19] In this regard, it specifically condemned the terms of access to the Celera human genome sequence data as "not consistent with the principles laid out in this report," noting that it permitted only "static access" for purposes of validation and not "dynamic access" for use in further research.[20]

Another important focus of debate has been the timing of data disclosure. The traditional trigger for data sharing in academic research is publication of research results. Large data sets, though, may not be ripe for publication in a prestigious journal until long after they are generated. Thus, research projects that aim to create large data sets over an extended period of time have presented special challenges for the implementation of data sharing norms.

In the genomics context, a series of international collaborative research efforts to create community resources for widespread use have prescribed data sharing policies that call for disclosure prior to publication.[21] In addition to facilitating prompt access to data for use in subsequent research, some of these efforts have also aimed to defeat corresponding patents, including patents on downstream inventions resulting from the data.[22] Within genomics, public research sponsors like NIH and the U.K.'s Wellcome Trust have applied normative pressure to achieve widespread data dissemination.

Outside the context of genomics, NIH has sought to use its leverage as a research sponsor to guide the data sharing practices of its grantees.[23] In recent years NIH has required researchers applying for more than

---

18. SHARING DATA & MATERIALS, *supra* note 6, at 5.

19. *Id.* at 34.

20. *Id.* at 48 box 3-2.

21. *See, e.g.*, The Human Genome Program of the U.S. Department of Energy Office of Science, Summary of Principles Agreed at the First International Strategy Meeting on Human Genome Sequencing—Bermuda (Feb. 25-28, 1996), http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml#1; WELLCOME TRUST, SHARING DATA FROM LARGE-SCALE BIOLOGICAL RESEARCH PROJECTS: A SYSTEM OF TRIPARTITE RESPONSIBILITY (2003), *available at* http://www.wellcome.ac.uk/assets/wtd003207.pdf [hereinafter TRIPARTITE RESPONSIBILITY].

22. *See* International HapMap Project, Genotype Access Registration, http://www.hapmap.org/cgi-perl/registration (last visited July 6, 2006).

23. *See* NIH, NOTICE NOT-OD-03-032, FINAL NIH STATEMENT ON SHARING RESEARCH DATA (2003), *available at* http://grants1.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html [hereinafter NIH STATEMENT].

$500,000 in funding to submit a plan for data sharing.[24] NIH cites a compelling list of arguments in support of such sharing, including reinforcing open scientific inquiry, facilitating new research, encouraging diversity of analysis and opinions, enabling the exploration of topics not envisioned by the original investigators, and permitting the creation of new data sets that combine data from different sources. The policy stops short of mandating data sharing, however, acknowledging the competing interest of "protecting confidential and proprietary data."[25]

While these international and federal initiatives provide useful benchmarks for CIRM to consider in formulating its own approach to data sharing, they do not constrain CIRM. In the patent context, the pervasive influence of the Bayh-Dole Act on publicly-sponsored research institutions is likely to constrain even a relatively large state-sponsored research initiative such as CIRM. These institutions actively seek and already hold many patents on stem cell technology.[26] By contrast, intellectual property rights for data are less clearly defined and institutional practices are less standardized. Given the variability in approaches to data sharing within the biomedical research community, CIRM may be well-positioned by virtue of the scale of its operation and the scarcity of federal funding for stem cell research to take a leadership role in setting the terms for data sharing in this context.

This Article discusses data sharing in California's stem cell initiative against the background of other data sharing efforts and in light of the competing interests that CIRM is directed to balance.[27] We begin by considering how IP law affects data sharing. We then assess the strategic considerations that guide the IP and data policies and strategies of federal, state, and private research sponsors. With this background, we discuss four specific sets of issues that public sponsors of data-rich research, including CIRM, are likely to confront: (1) how to motivate researchers to contribute data; (2) who should have access to the data and on what condi-

---

24. NIH, NIH DATA SHARING POLICY AND IMPLEMENTATION GUIDANCE (2003), *available at* http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm [hereinafter NIH DATA SHARING POLICY].

25. *Id.*

26. The most significant university patents are held by the Wisconsin Alumni Research Foundation (WARF). WARF holds broad patents on both embryonic stem cell lines in general and human embryonic stem cell lines in particular.

27. For purposes of this Article's analysis, we take these interests as a given. Thus, we do not evaluate, for example, whether CIRM's interest in providing financial benefit to the State of California is appropriate. Rather, we confine our analysis to possible conflict between the various CIRM interests.

tions; (3) what data get deposited and when; and (4) how to establish database architecture and curate and maintain the database.

II. THE ROLE OF INTELLECTUAL PROPERTY LAW IN DATA SHARING

Neither copyright nor patent law offers federal statutory protection for data as such. Indeed, both copyright law and patent law treat the informational content of writings and inventions as a spillover benefit for the public, while limiting the exclusionary rights of creators to something else: an original expression in the case of copyright,[28] and a product or process in the case of patent.[29]

On one reading, the failure to protect information under either patent or copyright law suggests that information gets no respect. This is the sense that emerges from reading copyright cases like *Feist Publications, Inc. v. Rural Telephone Services Co.*,[30] in which the Supreme Court re-

28. *Cf.* Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991) (holding that an alphabetized list of names and phone numbers lacked the minimum originality necessary for copyright protection, even though considerable effort may have gone into creating it).

29. Patentable subject matter is limited by statute to any new and useful process, machine, manufacture, or composition of matter, 35 U.S.C. § 101 (2000), all generally understood to be distinct from data or information. The subject matter boundaries of the patent system have been diminishing in recent judicial decisions in the face of creative claiming strategies for new technologies, particularly information technology. *See, e.g.*, State Street Bank & Trust Co. v. Signature Fin. Group, 149 F.3d 1368 (Fed. Cir. 1998), *cert. denied*, 525 U.S. 1093 (1999); AT&T Corp. v. Excel Commc'ns, Inc., 172 F.3d 1352 (Fed. Cir. 1999). Last term, the Supreme Court granted certiorari in the case of *Laboratory Corp. of America Holdings v. Metabolite Laboratories, Inc.*, 370 F.3d 1354 (Fed. Cir. 2004), *cert. granted*, 126 S. Ct. 543 (2005), *vacated, reconsidered, and cert. granted*, 126 S. Ct. 601 (2005), limiting the scope of its review to the question of patentable subject matter. Ultimately the Court dismissed the writ of certiorari as improvidently granted, with three justices dissenting. *Lab. Corp. of Am. Holdings v. Metabolite Labs., Inc.*, 126 S. Ct. 2976, 2921 (Breyer, J. with whom Stevens, J. and Souter, J. join, dissenting). Although the case ultimately failed to generate an authoritative opinion from a majority of the Supreme Court, the numerous amicus briefs filed in support of the defendant suggest a surprising level of discomfort in the business community with the trend toward more expansive patent eligibility. *See* Rebecca S. Eisenberg, *Biotech Patents: Looking Backward While Moving Forward*, 24 NATURE BIOTECH. 317 (2006).

30. 499 U.S. 340 (1991). The *sine qua non* of copyright is originality. To qualify for copyright protection, a work must be original to the author. Original, as the term is used in copyright, means only that the work was independently created by the author (as opposed to copied from other works), and that it possesses at least some minimal degree of creativity. To be sure, the requisite level of creativity is extremely low; even a slight amount will suffice. The vast majority of works make the grade quite easily, as they pos-

jected a claim of copyright in an alphabetized list of names and phone numbers. In this story, copyright law treats information as a mere byproduct of efforts that deserve protection only insofar as they yield some other, more creative output. Contemporary critics charge that copyright law has failed to appreciate the importance of information as an artifact of human ingenuity with value in its own right. In this view, as this value grows and becomes more vulnerable to misappropriation with the expanding capabilities of IT, this limitation on legal rights becomes more anomalous.[31]

From another perspective, the failure to protect data may reflect a reverence for information. Information is so valuable that society will not permit it to be monopolized. This is the sense that emerges from reading cases about disclosure in the patent system, in which courts treat the informational content of patent applications as the public's quid pro quo that justifies the issuance of patents.[32] In this story, disclosure of unprotected information is not an incidental byproduct of a process that aims to motivate something more worthwhile, but is the whole purpose of the system. We promote disclosure of precious information by rewarding disclosure with exclusionary rights in something else.

By requiring public disclosure of information about an invention while limiting the exclusive rights to the inventions defined in claims, patent law

---

sess some creative spark, "'no matter how crude, humble or obvious' it might be . . . [F]acts do not owe their origin to an act of authorship. The distinction is one between creation and discovery: the first person to find and report a particular fact has not created the fact; he or she has merely discovered its existence . . . [O]ne who discovers a fact is not its 'maker' or 'originator.' 'The discoverer merely finds and records.'" *Id.* at 345-47 (citations omitted).

31. *See, e.g.*, Jane C. Ginsburg, *U.S. Initiatives to Protect Works of Low Authorship*, *in* EXPANDING THE BOUNDARIES OF INTELLECTUAL PROPERTY: INNOVATION POLICY FOR THE KNOWLEDGE SOCIETY (R. Dreyfuss et al. eds., 2001).

32. *See, e.g.*, Kewanee Oil Co. v. Bicron Corp., 416 U.S. 470, 481 (1974) ("When a patent is granted and the information contained in it is circulated to the general public and those especially skilled in the trade, such additions to the general store of knowledge are of such importance to the public weal that the Federal Government is willing to pay the high price of 17 years of exclusive use for its disclosure, which disclosure, it is assumed, will stimulate ideas and the eventual development of further significant advances in the art."); Bonito Boats, Inc. v. Thunder Craft Boats, Inc., 489 U.S. 141, 151 (1989) ("[T]he ultimate goal of the patent system is to bring new ideas and technologies into the public domain through disclosure. State law protection for ideas and designs whose disclosure has already been induced by market rewards may conflict with the very purpose of the patent laws by decreasing the range of ideas available as the building blocks of further innovation."); United States v. Dubilier Condenser Corp., 289 U.S. 178, 186 (1933) ("[The inventor] may keep his invention secret and reap its fruits indefinitely. In consideration of its disclosure and the consequent benefit to the community, the patent is granted.").

not only fails to protect information but actually pushes it into the public domain as a spillover.[33] Yet, while the information disclosed in a patent application is publicly available, the exclusionary rights from the patent might still protect the patent owner from its unauthorized use if the use involves infringing the patent claims. If an inventor discloses in a patent application how to make and use a new mousetrap and a patent issues with claims drawn to the mousetrap, anyone who follows the directions in the disclosure to make and use the claimed mousetrap would be liable for infringement. A reader, on the other hand, who uses the disclosed information to problem-solve and devise a new spring-loaded device falling outside the scope of the mousetrap patent claims would not be liable, though the patent disclosure may have been invaluable to the reader in solving his problem. While patent claims legally constrain the use of information disclosed in patent specifications, the public disclosure of the information may also facilitate other non-infringing uses of that information.

Patent law concerning the scope of "prior art" that is used to evaluate the patentability of inventions has complex effects on incentives for information disclosure. The rules of patentability count all publicly available information, including the inventor's own disclosures, as prior art.[34] Consequently, those who hope to file patent applications may be inclined to defer disclosure of data until after filing related patent applications. On the other hand, those who wish to defeat the potential patent applications of their scientific or commercial rivals may disclose information early in the hope of creating more prior art.[35] The creation of patent-defeating prior art

33. *See* 35 U.S.C. § 112 ("The specification shall contain a written description of the invention, and of the manner and process of making and using it, in such full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains, or with which it is most nearly connected, to make and use the same, and shall set forth the best mode contemplated by the inventor of carrying out his invention."). An inventor who fails to file a patent application within a year of putting an invention to use loses the right to obtain a U.S. patent, *id.* § 102(b), forcing inventors to choose promptly between entering into the bargain of disclosure in exchange for a patent or secrecy and loss of right to patent.

34. *Id.* §§ 102-103. An inventor's own disclosures will not defeat the novelty of an invention under U.S. law because they do not show prior invention, knowledge, or use by another prior to the invention date, *id.* § 102(a), (g), but they may nonetheless give rise to a "statutory bar" against a patent if the disclosure occurred more than a year before the inventor's filing date. *Id.* § 102(b).

35. *See* Gideon Parchomovsky, *Publish or Perish*, 98 MICH. L. REV. 926 (2000); Douglas Lichtman, Kate Kraus & Scott Baker, *Strategic Disclosure in the Patent System*, 53 VAND. L. REV. 2175 (2000); Rebecca S. Eisenberg, *The Promise and Perils of Strategic Publication to Create Prior Art: A Response to Professor Parchomovsky*, 98 MICH. L. REV. 2358 (2000).

appears to have played a role in the development of disclosure rules for some large-scale biological resource projects.[36]

III. STRATEGIC CONSIDERATIONS OF SPONSORS IN DATA SHARING

A. Private Sponsors

Absent statutory protection, such as a patent or copyright, that survives beyond disclosure, a standard commercial strategy for preserving the value of data and databases has been secrecy, or more accurately, restricted access. Some owners of valuable databases permit only internal access to the data. Others make data available only to paying subscribers under the terms of database access agreements. Such owners may protect their databases as trade secrets, or at a minimum, under the law of contracts. Even without having to enforce legal rights in court, database owners may exercise considerable practical control over data sharing by restricting online access to databases to only particular internet addresses.

These strategies allow database owners to exclude free riders, and perhaps thereby capture enough value to justify creating the database. All the same, they are wasteful from a social perspective. These strategies restrict the dissemination of information that would have greater social value if more widely used and that could be made freely available at minimal cost. Restricting access leads to socially wasteful duplication as competitors create similar databases for their own use. It encumbers data consolidation, making it more difficult to aggregate data from multiple sources to create more comprehensive databases. Nonetheless, trade secrecy, contracts, and digital technology have an important role to play in encouraging firms to invest in the creation of databases.

B. Public Sponsors

The case for trade secrecy and other measures is weaker for information generated at public expense. Public funding mitigates concerns about the adequacy of incentives to generate information and makes the social waste inherent in secrecy more troubling. While some value may be created by interactions between creators and users of data when creators control access to data, broad dissemination often better serves the mission of public sponsors to advance science.[37] Further, data disclosure can provide

---

36 *See infra* Part IV.A. and note 79.

37. *Compare* Ashish Arora & Robert Merges, *Specialized Supply Firms, Property Rights, and Firm Boundaries*, 13 INDUS. & CORP. CHANGE 451 (2004) (discussing value

a valuable check on fraudulent research claims. This risk has, regrettably, become salient in the recent experience of stem cell research.[38] Data disclosure also provides a check against over-claiming in the political arena, another concern for stem cell research.[39]

Public sponsors have an interest not only in advancing science but also in ensuring that research discoveries made in the course of funded research are effectively disseminated and practically utilized. The Bayh-Dole Act emphasizes this interest and aims to promote it by encouraging grantees to patent their inventions and then to license these patents to firms that will undertake further development and commercialization.[40] The theory is that licenses, especially exclusive licenses, will provide necessary protection against competition during the risky and costly commercialization process. Although one might expect the interests of state sponsors to be similar to those of the federal government, CIRM in fact faces more significant (and more parochial) constraints under the terms of Proposition 71.

---

of customization of research inputs for particular users), *with* NIH STATEMENT, *supra* note 23.

38. *See* Sei Chong & Dennis Normile, *How Young Korean Researchers Helped Unearth a Scandal. . . And How the Problems Eluded Peer Reviewers and Editors*, 311 SCI. 22-25 (2006).

39. *See* David A. Shaywitz, *Stem Cell Hype and Hope*, WASH. POST, Jan. 12, 2006, at A21.

40. 35 U.S.C. § 200. Other interests noted in the Bayh-Dole statute include encouraging participation of small business firms in federally supported research and development (R&D), promoting collaboration between commercial concerns and nonprofit organizations, promoting competition and enterprise without unduly encumbering future R&D, promoting "the commercialization and public availability of inventions made in the United States by United States industry and labor," ensuring that the government obtains sufficient rights in federally supported inventions to meet its needs, and minimizing administrative costs. *Id.* Federal research sponsors are not charged by statute with recovering revenues from technologies patented by grantees except in the case of inventions made in a government-owned, contractor-operated facility (i.e. a national laboratory). Under 35 U.S.C. § 202(c)(7), sponsors are directed to include in funding agreements requirements for sharing royalties with inventors and for using remaining income, after payment of costs, to support scientific research or education. A different rule applies to funding agreements for the operation of a government-owned, contractor-operated facility; these agreements are to require payment to the U.S. Treasury of 75% of the excess revenues after payment of expenses if the balance exceeds 5% of the annual budget of the facility. *Id.* § 202(c)(7)(E). Although the Bayh-Dole Act directs grantees to give a preference in the award of exclusive licenses to firms that agree to manufacture the invention in the United States, if that constraint proves to be problematic, then the sponsor may waive it. *Id.* § 204.

In addition to promoting the development of stem cell therapies, Proposition 71 identifies a number of goals that are more narrowly focused on the interests of California constituencies, including: to "[p]rotect and benefit the California budget . . . by providing an opportunity for the state to benefit from royalties, patents, and licensing fees that result from the research"; to "[b]enefit the California economy by creating projects, jobs, and therapies that will generate millions of dollars in new tax revenues in our state"; and to "[a]dvance the biotech industry in California to world leadership, as an economic engine for California's future."[41] Proposition 71 enhances the likelihood that the California focus of these goals will be taken to heart by requiring California institutional affiliations for each member of the ICOC, the committee charged with governing CIRM.[42]

Of course, it is not at all surprising that a California voter initiative that appropriates $3 billion in research funding would promote the interests of California constituencies. Indeed, in the Bayh-Dole Act, the federal government made a similar move to promote the interests of U.S. firms by directing recipients of U.S. research funding to give preferences for exclusive licenses to firms that would manufacture in the U.S.[43] These strategies allow taxpayers to capture more of the benefits of tax-funded programs. To the extent that spillovers to non-local interests limit incentives for governments to invest in research and development ("R&D"), such strategies may be necessary to encourage government-funded R&D.

Nonetheless, state-focused preferences in the management of intellectual property are more limiting than national preferences, and thus are more troubling. If state-sponsored R&D initiatives become more prevalent, a proliferation of local preferences could threaten to balkanize valuable IP among the states, making it difficult for firms to collect the rights needed to move forward with product development. Even a single state-sponsored research initiative such as CIRM could significantly restrict dissemination through local preferences if it controls access to broad, cross-cutting technologies, like stem cells, that may have implications for a range of problems.[44]

---

41. *Proposition 71*, *supra* note 4, at 147.

42. *Id.* at 147-48; California Stem Cell Research and Cures Act, CAL. HEALTH & SAFETY CODE § 125290.20(a) (2006).

43. Note, though, that if that constraint proves to be problematic, the sponsor may waive it. 35 U.S.C. § 204.

44. It is interesting to compare the interests of state research sponsors in furthering the interests of local constituents with the interests of private research sponsors in furthering the interests of shareholders. Private sponsors are unlikely to care whether the money

Moreover, in contrast to the Bayh-Dole Act, Proposition 71 directs CIRM to recoup revenues for the California state treasury.[45] This revenue goal is in tension not only with the goal of ensuring widespread dissemination of research results, but also, to a lesser degree, with the goal of commercialization. To the extent that product developers are expected to return money to the state treasury, such a requirement acts as a tax on commercialization.

Although the Bayh-Dole Act and Proposition 71 focus on patent rights in technologies emerging from sponsored research, data sharing in the context of sponsored research poses similar tradeoffs between capturing value for political constituencies and promoting scientific progress.

IV. SPECIFIC CHALLENGES FOR CIRM

The challenge for CIRM is to capture an adequate return for its constituents on its investment in stem cell research without unduly limiting its overall social value. In examining this challenge, we address four highly interdependent issues that any effort to promote data sharing must consider:[46] (1) incentives to contribute data; (2) who gets access and under what conditions; (3) what gets deposited and when; and (4) database architecture, maintenance, and curation. Throughout our discussion, we draw upon the experiences of prior database initiatives, particularly those at the federal level, which have attempted to promote widespread dissemination and sharing. In the absence of information on the specific research CIRM is likely to fund, we make these observations at a relatively high level of generality.

A. Incentives to Contribute Data

In order to be effective, data release policies must give scientists clear incentives to contribute their data. This Section focuses on incentives in

---

they are making emanates from activity in California or in Massachusetts, and are therefore less likely to restrict dissemination on the basis of geography. On the other hand, private sponsors may be less likely than state counterparts to disseminate information in ways that benefit the public but do not benefit their own bottom lines. CIRM might be content to spend money in ways that mean more medical treatments and more jobs for California voters even if no money flows back to the state coffers, but commercial firms that are obligated to return value to shareholders cannot afford to be so public-spirited.

45. *Proposition 71*, *supra* note 4, at 147.

46. For purposes of this article, we put to one side thorny problems regarding privacy that might be raised by data associated with personally identifiable information. We will assume that data involved in stem cell research would not trigger concerns about personally identifiable information or that the data could be effectively de-identified to address such concerns.

two somewhat distinct contexts: centralized data production projects and more decentralized, investigator-driven science.

As a general matter, incentives are necessary because most rewards in research science, including academic appointments, promotion, and grant funding, depend on a record of frequent publication. Scientists may perceive sharing data, even after an initial publication, as providing advantages to competitors in the race to generate further publications. Scientists may also be reluctant to share data because of involvement in commercial activities. Sharing may imperil patent applications or destroy trade secrecy. Emerging evidence reveals that some research communities in the life sciences are reluctant to share data even after publication. For example, a survey conducted by Eric Campbell and his colleagues found that 47% of academic geneticists who had made a request to another academic had been denied access to data or materials associated with a published article at least once in the preceding three years.[47] Scientific competition and commercial involvement were both important predictors of refusal to share.[48]

Although NIH now requires grant applicants to include a data sharing plan in grant applications exceeding $500,000 per year,[49] so far it has done little to enforce compliance. If CIRM wants its grantees to share data, it should consider mechanisms for ensuring compliance from the outset in order to offset the powerful incentives that scientists face to withhold access to data. Mechanisms might include rewards for compliance or sanctions for noncompliance, such as loss of continued funding. A possible reward might involve privileged access to data analysis tools for those who contribute data to an archive. CIRM could also track downloads of

47. Eric Campbell et al., *Data Withholding in Academic Genetics: Evidence from a National Survey*, 287 JAMA 473, 477 (2002). The Campbell study did not distinguish between data and tangible materials. Because one important impediment to sharing identified by the study—the effort and financial cost associated with replication and transfer, *id.* at 478, —is much lower for data than for tangible materials, the study may overestimate impediments to data disclosure. *Cf.* John Walsh et al., *View from the Bench: Patents and Materials Transfers*, 309 SCI. 2002 (2005) (indicating that problems in transfer of tangible materials appear to have risen since Campbell's study, but not addressing the question of data). Nonetheless, as discussed earlier, a series of workshops and reports emanating from the biomedical research community confirms a growing perception of departures from the principle of data sharing upon publication.

48. *See* Campbell, *supra* note 47, at 478.

49. *See* NIH DATA SHARING POLICY, *supra* note 24.

data from a centralized archive and give special acknowledgements or other rewards to scientists whose data was downloaded frequently.[50]

It may be easier to achieve compliance with a data sharing plan within a tightly knit community of scientists. For example, at the height of the Human Genome Project (HGP), the five major production labs that contributed large amounts of sequence to the public GenBank database teleconferenced on a weekly basis.[51] In this environment the normative pressure to comply with data disclosure—even pre-publication disclosure—was unusually strong. Some data users from the HGP and other community resource projects have also argued that widespread data availability was the quid pro quo for the major centers receiving large sums of money to complete these projects without undergoing peer review of each individual portion.[52] CIRM may be able to create similar normative pressure to comply with data disclosure obligations if it funds large-scale, centralized data production.[53]

It bears emphasis, though, that researchers in the HGP were motivated not only by a public-spirited desire to make data quickly available (without any background patents on associated material)[54] but also by a competitive desire to outdo rival private sector efforts. Measures of the volume of data accumulating in GenBank served as a conspicuous marker of accelerating productivity for the HGP. Public availability served as a salient point of distinction from the proprietary databases of commercial rivals.

---

50. Although rewards of this sort might not be as attractive as preserving exclusive access so as to mine the data for additional publications (particularly if university tenure and promotion committees continue their current practice of considering publication to be the primary benchmark of success), they might provide some incentive.

51. JOHN SULSTON & GEORGINA FERRY, THE COMMON THREAD: A STORY OF SCIENCE, POLITICS, ETHICS AND THE HUMAN GENOME 193 (2002) (discussing the Friday conference calls that took place among the "G5" to coordinate activities).

52. Steven Salzberg et al., *Unrestricted Free Access Works and Must Continue*, 422 NATURE 801 (2003) (correspondence from bioinformaticians arguing that obligations of scientists in large-scale data production centers differ from those of traditional scientists).

53. For example, CIRM might fund a group of centers to produce data on gene expression at different stages of stem cell differentiation.

54. In February 1996, scientists from the major sequencing centers in the HGP explicitly disavowed patenting. Eliot Marsh, *Data Sharing: Genome Researchers Take the Pledge*, 272 SCI. 477, 477 (1996). NIH followed up with an April 1996 policy statement strongly discouraging patenting by HGP grantees. National Human Genome Research Institute, NHGRI Policy Regarding Intellectual Property of Human Genomic Sequence—Apr. 9, 1996, http://www.genome.gov/10000926. Though it may be in some tension with Bayh-Dole, this "no patenting norm" has also been part of subsequent NIH-sponsored "community resource" projects. *See* Arti K. Rai & Rebecca S. Eisenberg, *The Public Domain: Bayh-Dole Reform and the Progress of Biomedicine*, 66 LAW & CONTEMP. PROBS. 289 (2003).

Public access helped to justify continued public support for a project that appeared to duplicate work being done in the private sector. Moreover, rapid data availability might have been expected to frustrate commercial rivals by creating prior art to defeat future gene sequence patents.[55] Rapid public disclosure also undermined the viability of private sector business models that entailed charging license fees for database access. Although they were able to raise investment capital to create their databases, private sector rivals were ultimately not able to survive in the database business.[56]

Given its mandate to "[a]dvance the biotech industry in California to world leadership, as an economic engine for California's future,"[57] it seems unlikely that CIRM would want to drive out private sector data producers in any large-scale data production efforts that it might fund. CIRM might, therefore, count impediments to private R&D as a cost to weigh against the benefits of a public domain approach to research inputs like data. A public domain approach eliminates the significant costs that are likely to be associated with negotiating access, but it also imposes some costs of its own. In addition to making public funding necessary in many cases, aggressive versions of a public domain approach may undermine the types of small firms that tend to provide specialized research inputs in the marketplace. To the extent that these foregone market incentives for innovation by specialized firms are superior to the incentives that operate

55. Although raw genomic data would not undermine claims to specific genes of identified function, annotated data might do so. A major goal of annotation is to identify coding regions in the genome and add information about the function of the protein for which the region codes. A recent empirical study suggests that at least 20% of human genes are in fact covered by patents; some genes are covered by multiple patents. *See* Kyle Jensen & Fiona Murray, *Intellectual Property Landscape of the Human Genome*, 310 SCI. 239 (2005). The extent to which these patents are valid over the prior art is unclear.

56. The major private sector rival to the public database, Celera Genomics led by Craig Venter, was ultimately unsuccessful in its efforts to charge for its database and released its data into the public domain. Emma Marris, *Free Genome Databases Finally Defeat Celera*, 435 NATURE 6 (2005). Although public availability of the human genome avoids the potentially crippling costs that might have been associated with negotiating access, and is thus a welcome development, the presence of a private sector rival had some benefit. The private sector effort arguably provided the competition necessary for the public sector to work efficiently. In particular, private sector competition may have been the catalyst necessary to overcome the public sector's resistance to the whole genome shotgun sequencing approach, a methodology that has proved to be successful. *See* Rebecca S. Eisenberg & Richard Nelson, *Public vs. Proprietary Science: A Fruitful Tension?*, 131 DAEDALUS 89 (2002).

57. *Proposition 71*, *supra* note 4, at 147.

in large, vertically integrated firms or in the public sector,[58] that cost may be significant.

Unlike the HGP, most community resource projects in genomics have not sought to drive out private sector competitors. These projects may therefore provide a more appropriate model for CIRM. Non-HGP community resource projects have, of course, lacked the incentives for disclosure provided by a race with a high-profile private sector competitor. They have made up for the absence of such incentives, however, by explicitly seeking to preserve some of the rewards of publication for scientists who contribute to public databases prior to publication. A report from the Wellcome Trust on *Sharing Data from Large-Scale Biological Research Projects: A System of Tripartite Responsibility* proposes that producers of database resources publish a project description at the beginning of the project describing their plans.[59] These project descriptions should provide for production, analysis, and release of the data and give a citation for referencing the sources of the data.[60] The Wellcome Trust report not only admonishes data users to cite the proper reference source but it also urges them to "recognize that the resource producers have a legitimate interest in publishing prominent peer-reviewed reports describing and analyzing the resource that they have produced . . . ." Indeed, the report indicates that data users might best "promote the highest standards of respect for the scientific contribution of others," by discussing or coordinating their publication plans with resource producers.[61] In comparable community resource projects, CIRM could use its leverage with both data producers and the data users it funds to encourage compliance with these suggested principles.

There is an obvious tension between preserving opportunities for those who disclose data to publish their own future analyses and allowing outside users full access to the data. Unlike the Wellcome Trust report, which does not endorse explicit delays on publication by outside data users, some community resource projects have tried to restrict publication. One example is the Genetic Association Information Network (GAIN),[62] a public-private partnership of the NIH and several private firms, currently Pfizer,

58. *See* Arora & Merges, *supra* note 37; Ashish Arora et al., *Markets for Technology and Their Implications for Corporate Strategy*, 10 INDUS. & CORP. CHANGE 419 (2001).

59. *See* TRIPARTITE RESPONSIBILITY, *supra* note 21.

60. *Id.* at 3-4.

61. *Id.* at 4.

62. Foundation for the NIH, GAIN Program Home Page, http://www.fnih.org/GAIN/GAIN_home.shtml (last visited Aug. 3, 2006).

Affymetrix, and Abbott Laboratories.[63] GAIN aims to understand the complex set of genetic factors influencing risk for common diseases by conducting a series of whole genome association studies that employ samples from patients with such diseases.[64] The GAIN publication policy gives contributing investigators a period of nine months during which they have the exclusive right to submit publications based on their data.[65] At the same time, the policy gives approved users, who sign a restrictive agreement, access to the data during this period.[66] CIRM may need to consider whether the type of formal restriction on publication adopted by GAIN unduly favors initial producers of data relative to subsequent users.

In any event, the model adopted for community resource projects in genomics is likely to be inappropriate for decentralized, investigator-initiated work. Detailed information about the characteristics of all available stem cell lines, for example, is likely to emerge not from a top-down data production effort, but rather from decentralized contributions of individual labs. Stem cell scientists would presumably generate such information as they worked with, and published on, particular lines. A database that accumulated such information, which some stem cell scientists have proposed,[67] might include details of derivation, genetic details, and results indicating pluripotence and antibody markers.

For such work, the federally funded Protein Data Bank (PDB) may be a better model. In 1971, a group of crystallographers established the PDB as a centralized repository for three-dimensional protein structure data. Deposit of structures, though, did not begin in earnest until the 1980s, as the community began to see collective advantages of deposition. In 1989, the International Union of Crystallography (IUCr) reinforced community views by calling on researchers to deposit data once they submitted for publication a research article based on the data.[68] Actual data release,

---

63. *See* Foundation for the NIH, GAIN Program Partnerships, http://www.fnih.org/GAIN/Partnerships.shtml (last visited Aug. 3, 2006).

64. *See* Foundation for the NIH, GAIN Program Overview, http://www.fnih.org/GAIN/Background.shtml#Program (last visited Aug. 3, 2006).

65. Foundation for the NIH, Policies and Procedures: GAIN Publication Policy, http://www.fnih.org/GAIN/policies.shtml#Publication (last visited Aug. 3, 2006) [hereinafter GAIN Publication Policy]; Foundation for the NIH, GAIN Data Use Certification Terms of Access, http://www.fnih.org/GAIN/documents/Data_Use_Certification.pdf, ¶ 6 (last visited Aug. 3, 2006) [hereinafter GAIN Terms of Access].

66. GAIN Publication Policy, *supra* note 65.

67. Krishanu Saha, Navigating to the Right Stem Cell Line (working paper, on file with authors). A preliminary version of such a database is currently available at The Stem Cell Community, http://www.stemcellcommunity.org (last visited Aug. 3, 2006).

68. SHARING DATA & MATERIALS, *supra* note 6, at 74-75.

however, did not have to be immediate: the IUCr allowed researchers to request a one-year hold before public release of the data by the database.[69] IUCr justified this one-year hold as a reward for the difficulties in determining protein structure. As these difficulties decreased, leaders within the community began to call for immediate release of data upon publication. In 1999, the NIH announced a policy of data release upon publication for its grantees.[70] Major scientific journals such as *Science* and *Nature* now require data deposition in PDB as a condition of publication.[71]

In contrast with recent community resource projects in genomics, the PDB effort does not have a prohibition on patenting. Although the PDB does not keep track of background patents,[72] protein structure data could be associated with background patents on the gene, protein crystal, or perhaps even on a computer model of a protein binding pocket that purports to allow the investigator to test drug candidates.[73] In a decentralized project such as PDB, a prohibition on patents might have served as a significant disincentive to scientific participation.

The PDB story exemplifies cooperation between scientific leaders in the protein crystallographic community and research sponsors over several decades to make data deposition an essential aspect of publication.[74] A similar combination of sustained sponsor pressure and leadership from key leaders in the stem cell community may also be critical in order for data sharing in routine CIRM-funded work to succeed.

B. Access: By Whom and Under What Conditions

Incentives to contribute are also likely to be affected by scientists' perceptions regarding who may access their contributions, and under what conditions. The issue of access is an important one, both for ensuring maximum benefit from CIRM-sponsored research and for determining how CIRM, and the state of California more generally, reap returns on their investment.

A pure public domain approach to scientific resources would place no restrictions on who could seek access or on what they could seek. In the

---

69. *Id.* at 75.

70. *Id.* at 76.

71. *See* Science, Database Deposition Policy, http://www.sciencemag.org/about/authors/prep/gen_info.dtl#datadep (last visited Aug. 3, 2006).

72. Telephone Interview with Helen Berman, Professor, Department of Chemistry and Chemical Biology, Rutgers University, in Piscataway, N.J. (Mar. 2, 2005). Professor Berman is a leader of the PDB community.

73. Although the last category of patent appears quite close to a patent on data, the U.S. Patent & Trademark Office has issued such patents.

74. Interview with Helen Berman, *supra* note 72.

area of publication-related biomedical materials, CIRM has already departed from a pure public domain approach in favor of a policy that favors California researchers. The CIRM IPPNPO requires grantees to share biomedical materials described in published scientific articles within 60 days of receiving a request for such materials. But the IPPNPO appears to limit grantee obligations to those who are seeking the materials for "research purposes in California."[75]

CIRM might similarly choose a tiered approach to data access in order to benefit various constituents. It might, for example, permit access by: (1) CIRM-funded nonprofit researchers only; (2) all CIRM-funded researchers; (3) all California researchers; (4) all stem cell researchers who had contributed their own data (and/or agreed to contribute their own annotations/improvements to the database); or (5) all stem cell researchers. Certain categories of researchers could be excluded altogether or could be given access under restrictive conditions. CIRM could require for-profit institutions, or non-California institutions, to pay for access. Non-price methods of tiering, such as early access by certain favored categories of researchers, could favor preferred groups while still permitting broad access.

Providing preferential access to CIRM-funded researchers, or to researchers based in California, could promote Proposition 71's goal of stimulating the California economy. Charging for-profit institutions for access may promote its goal of direct returns to the California budget. Furthermore, giving preference to those who themselves contribute data, whether through initial contributions or through improvements or annotations to the initial contribution, could provide an additional incentive to contribute.

These benefits come at some cost though: the more conditions CIRM places on access, the more potential investigators are excluded. Moreover, because data are not protected by intellectual property rights, contract-based access must specifically include restrictions against the possibility of dissemination to third parties. Thus, in order for any contractual restrictions to be effective, they must include a restriction on further dissemination.

Again, recent experience with publicly funded genomics databases provides a useful background for examining the costs and benefits of restricting access. In the case of the HGP, data were released into the public

75. IPPNPO, *supra* note 1, at 16. Similarly, the IPPNPO restricts its requirement that CIRM-funded patents materials be made available for research purposes to "California research institutions." *Id.*

domain without restriction. The public domain approach was chosen over the objection of some public sector scientists who did not view creating prior art as the best weapon for defeating proprietary claims. Because the data were freely available, those who accessed the data could blend it with their own privately-held information and make the combination proprietary.[76] These scientists suspected that Craig Venter, the major private sector challenger to the HGP, had adopted this approach.[77]

The frustration of these public sector scientists appears to have influenced the approach toward data sharing in subsequent community resource projects. For example, the International Haplotype Map (HapMap) project, which receives funding from both the NIH and the Wellcome Trust, initially took a very different approach to data release. In that case, the raw data on single base DNA variations, also known as single nucleotide polymorphisms (SNPs), were not released into the public domain. Rather, they were made available via a clickwrap license explicitly modeled on the General Public License (GPL) used by open source software developers.[78] Until December 2004, when the license restrictions were lifted, the license prohibited licensees from combining the data with their own so as to seek product patents on combinations of SNPs known as haplotypes.[79]

The HapMap experience illustrates some of the difficulties involved in adapting the GPL to the release of biomedical research data.[80] First, the

---

76. SULSTON & FERRY, *supra* note 51, at 211-13.

77. There is some controversy over the extent to which the Venter project actually relied on the public data. *Compare* Robert H. Waterston et al., *More on the Sequencing of the Human Genome*, 100 PROC. NAT'L ACAD. SCI. 3022, 3024 (2003) (claiming that Celera's assembly is "appropriately viewed as a refinement built on the HGP assemblies") *with* Mark D. Adams et al., *The Independence of Our Genome Assemblies*, 100 PROC. NAT'L ACAD. SCI. 3025, 3026 (2003) (claiming that Celera produced an "independent assembly" and that HGP contribution to the structure and content was minimal).

78. International HapMap Project, Public Access License–Version 1.1, Aug. 2003, http://www.hapmap.org/cgi-perl/registration [hereinafter HapMap License]. The HapMap License includes an acknowledgement to the GNU General Public License of the Free Software Foundation. *Id.*

79. *See id.* ¶ 2(b)(i) ("[Y]ou shall not file any patent applications that contain claims to any composition of matter of any single nucleotide polymorphism ('SNP'), genotype or haplotype data obtained from the Genotype Database or any SNP, haplotype or haplotype block based on data obtained from the Genotype Database.") Haplotypes are SNP clusters that are inherited together. Haplotypes associated with particular phenotypes can be used as markers for diagnostic tests and drug targets. *See generally* International HapMap Project, What is the HapMap?, http://www.hapmap.org/whatishapmap.html.en (last visited Aug. 7, 2006).

80. For a general discussion of "open source" approaches in biomedical research, see Arti K. Rai, *Open and Collaborative Research: A New Model for Biomedicine, in*

GPL is structured as a license to intellectual property rights. In the context of open source software, the licensed rights consist of copyright in software, a right that has been recognized both by Congress and by the courts. Under U.S. law, there is no comparable intellectual property right in data to anchor the HapMap license. The HapMap license denies this difficulty, requiring those who would access the data to acknowledge, contrary to legal authority, that the data are protected by U.S. copyright law.[81]

Second, because there is no property right that survives disclosure to those not bound by the license, in order to ensure that third parties do not gain access to the data without agreeing to the terms of the license, the HapMap license imposes tight restrictions on dissemination. Researchers who accessed the data prior to December 2004 could not release the data to anyone who was not bound by the same license terms. Most notably, they could not include the data in publications based on the data.[82]

Third, the GPL is designed to preclude all downstream restrictions on dissemination, an approach that is possible in the area of software, where intellectual property has never been a particularly strong driver of R&D. In contrast, in the biopharmaceutical area, patents—particularly downstream patents on therapeutics—are clearly important. The HapMap license seeks to avoid imperiling downstream patents that might matter for future product development through the use of complex and ambiguous license provisions. These provisions appear to prohibit product patents on

---

INTELLECTUAL PROPERTY RIGHTS IN FRONTIER INDUSTRIES: SOFTWARE AND BIOTECH (Robert Hahn ed., 2005).

81. The license states in relevant part: "You acknowledge that the Genotype Database and the data contained in it, to which access is provided under the terms of this License, are protected by law including, but not limited to, copyright laws of the United States . . . ", HapMap License, *supra* note 78, ¶ 5.

82. International HapMap Project, Data Access Policy, http://www.hapmap.org/cgi-perl/registration, ¶ G [hereinafter HapMap Data Policy] ("[While] you are free to publish the results of those analyses [of genotypic information], you may not include in such publications the details of the individual genotypes that the Project has not yet released.").

SNPs or haplotypes[83] but may allow for claims to certain uses of SNPs and haplotypes.[84]

Finally, the enforceability of open source licenses remains a somewhat open question. Clickwrap licenses are generally considered enforceable contracts, so long as the licensee has had the opportunity to view and assent to the terms.[85] However, if a public funding agency were to bring a breach of contract action against a license violator, the measure of damages would be unclear. Perhaps alleged infringers of patents that were obtained or enforced in violation of the agreement could assert that the patents were invalid or unenforceable for inequitable conduct, but there is no clear authority for such an argument. It may be that such agreements are better understood as efforts to define norms of forbearance from enforcement of intellectual property rights within a scientific community than as binding agreements that are themselves enforceable in a court of law.

More recent community resource projects have been less aggressive in their approach to restricting future intellectual property claims. Like the HapMap license, the GAIN Data Use Certification requires those who access the data to refrain from disclosing the data to anyone who is not bound by the same agreement.[86] It also urges registrants not to rely on GAIN-supported data to seek patents on markers that might be useful in diagnosis or identification of drug targets.[87] However, the language is entirely hortatory, calling upon approved users to "acknowledge the intent" of the GAIN IP policy, reminding them that "[i]n this spirit, it is expected" that data and conclusions will remain freely available, and stating that GAIN "encourages" compliance with various NIH policies that favor shar-

83. HapMap License, *supra* note 78, ¶ 2(b)(i). The policy explaining the license is more ambiguous on the question of product patents. It suggests that patents, presumably both product and process patents, on haplotypes with identified utility are acceptable so long as they do not block access to the underlying HapMap Data. *See* HapMap Data Policy, *supra* note 82, ¶ E ("This licensing approach is not intended to block the ability of users to file for intellectual property protection on specific haplotypes for which they have identified associated phenotypes, such as disease susceptibility, drug responsiveness, or other biological utility, as long as public access to, and use of, the data produced by the HapMap Project is preserved.").

84. HapMap License, *supra* note 78, ¶ 2(b)(ii) ("[Y]ou shall not file any patent applications that contain claims to particular uses of any SNP, genotype or haplotype data obtained from the Genotype Database or any SNP, haplotype or haplotype block based on data obtained from, the Genotype Database, unless such claims do not restrict, or are licensed on such terms that they do not restrict, the ability of others to use at no cost the Genotype Database or the data that it contains for other purposes.").

85. *See, e.g.*, Davidson & Assocs. v. Jung, 422 F.3d 630 (8th Cir. 2005).

86. GAIN Terms of Access, *supra* note 65, ¶ 4.

87. *Id.* ¶ 5.

ing.[88] Further, the document explicitly "recognizes the importance of the later development of IP on downstream discoveries, especially in therapeutics."[89]

The less rigid language used in the GAIN Data Use Certification makes good sense given the difficulty of determining ex ante just which patents will prove necessary to preserve economic incentives for product development in the biopharmaceutical area.[90] A small diminution in the incentives of public sector database contributors to contribute their data is a price worth paying for a safeguard against destruction of future incentives for product development.

In sum, experience with restrictions on access to genomics databases suggests that contract-based restrictions on access can provide incentives for data producers to contribute their data. Indeed, data producers may strongly prefer such restrictions. Contractual restrictions, however, are very difficult to enforce without sacrificing dissemination. Contractual restrictions on future intellectual property rights may be particularly ill-advised in an area as sensitive to patents as biomedical science.

C. What Gets Deposited and When

A third set of questions concerns what data get deposited and when. One benchmark is the standard set in the National Research Council report *Sharing Publication-Related Data and Materials*. This report calls for disclosure of "whatever is necessary to support the major claims of the paper and would enable one skilled in the art to verify or replicate the claims."[91] Tying disclosure obligations to publication has implications for both the scope and timing of disclosure obligations.

With respect to the scope of disclosure, the focus on verification and replication of publication claims allows for evolution in standards of disclosure over time within a given scientific community. In the case of the Protein Data Bank, for example, requirements for what gets deposited have evolved. Initially, crystallographers only deposited atomic coordinates. However, scientists subsequently determined that atomic coordinates did not necessarily provide all the information necessary for verification and improvement. Today there is general agreement that structural

88. *Id.*

89. *Id.*

90. Alternatively, it may reflect a recognition that simple release of GAIN-supported data is all that is necessary to invalidate marker patents.

91. SHARING DATA & MATERIALS, *supra* note 6, at 5.

factors—the raw information from which researchers derive coordinates—should also be deposited.[92]

The issue of when data should be deposited is a critical one. As already noted, for community resource projects in genomics, the public sponsors have generally required immediate, *pre-publication* deposit.[93] CIRM should recognize, though, that pre-publication release of data is highly unusual in science. The data release policies for community resource projects in genomics offer a precedent for centralized data production projects that CIRM might fund. However, it is unlikely that scientists could be persuaded to agree to pre-publication release beyond that context. As discussed earlier, the current structure of investigator-driven academic science virtually requires some level of secrecy prior to publication. In this context, pre-publication data release might even be undesirable because it would interfere with the incentives provided by the reputation benefits attached to publication.

On the other hand, a significant drawback to the current system of tying data release to publication is that negative data often remain undisclosed. CIRM might be able to address this bias in a data release policy by requiring disclosure not only of the data that leads to the publication but also of any negative data that emerge along the way. Indeed, because negative data can prove highly useful for future researchers, CIRM would perform a valuable service by establishing data archives that require deposits of both positive and negative data.

If disclosure obligations are not tied to publication, it becomes necessary to establish another marker to signal when data are ripe for release. In the case of the HGP, the community originally determined that sequence assemblies of 1-2 kilobases or greater should be released. However, when the community switched in part to a different sequencing methodology that did not assemble completed sequences until much later in the project, it determined that tying data release to assembly was no longer appropriate. In 2000, NIH extended its release policy to include submission of raw sequence traces.[94]

Finally, it bears emphasis that the distinction between pre-publication data deposit and data deposit upon publication rests on a model that currently prevails in the life sciences in which peer review precedes print

---

92. Interview with Helen Berman, *supra* note 72.

93. National Human Genome Research Institute, Reaffirmation and Extension of NHGRI Rapid Data Release Policies: Large-Scale Sequencing and Other Community Resource Projects—February 2003, http://www.genome.gov/10506537.

94. *See id.*

publication. This distinction may become less important in the future if the life sciences community adopts a model similar to that used in the physics community, as well as in other scholarly communities, where Web-based publication precedes peer review.

In the near term, publication is likely to provide a useful benchmark for both the timing and scope of data disclosure for most CIRM-funded research. This approach is less likely to disrupt traditional scientific rewards and incentives than a system of pre-publication disclosure, making it easier to persuade scientists to comply. It has the further advantage of allowing CIRM to rely on the judgments of journal editors and peer reviewers in determining when research results are ripe for disclosure.

D. Database Architecture, Curation, and Maintenance

A last set of issues relates to database architecture, curation, and maintenance. Such issues tend to be neglected, but they are critical to the long-term survival and usefulness of databases.

A centralized, Web-based data archive is the most obvious platform for data sharing. In biomedical research, some of the most prominent databases—GenBank for DNA sequence data and the PDB for 3-D structure data—are centralized repositories. A major advantage of a centralized database is that data are prominently available in a uniform, readily searchable format. Disadvantages include cost and the need for agreement on data standards. Even with these disadvantages, a centralized database is probably most appropriate for data that are most useful when aggregated, such as data on gene expression or on the characteristics of available stem cell lines.

Another format that might prove useful for certain projects is a federated approach, in which data are maintained and controlled at the level of the individual lab but can be integrated across databases. Federated systems might be useful even in situations where the core data reside on a central computer or server. For example, the distributed annotation system (DAS) that can be used on genomic data deposited at EMBL, the European counterpart to Genbank, allows those who want to annotate genomic data to do so on their own servers. Other DAS users can then designate which server annotations to layer over the core data.[95]

---

95. Telephone Interview with Lincoln Stein, Researcher, Cold Spring Harbor Laboratory, in Cold Spring Harbor, N.Y. (May 13, 2005); *see also* Lincoln D. Stein, Sean Eddy & Robin Dowell, Distributed Sequence Annotation System, http://biodas.org/documents/rationale.html (last visited Aug. 26, 2006).

The format that is probably least useful, but may nonetheless be sufficient for certain investigator-initiated projects, is posting on a local lab server. This format maximizes investigator control over the data but is relatively inconvenient for access by other users.

For all three types of databases—centralized, federated, and local—funding for ongoing curation and maintenance is critical. Indeed, one of the central problems facing life sciences databases today is that funds for curation and maintenance are often not available. A recent survey of eighty-nine life science databases determined that fifty-one are struggling financially: they have either been shut down for lack of funding or are being updated sporadically.[96] As it considers what types of research to fund, CIRM should be aware of the importance of providing funding for the ongoing curation and maintenance of databases that serve as important resources for the stem cell community.

V. CONCLUSION

Proposition 71 calls upon CIRM to balance a number of competing interests, including not only scientific progress but also commercialization of research results and financial returns to the State of California. In the context of patenting, licensing, and tangible research materials, CIRM has enunciated a detailed plan for balancing these competing interests. With respect to the important issue of data sharing, however, the balance that CIRM aims to strike is less clear. Data sharing represents a significant opportunity for a show of leadership. The federal example binds CIRM less directly in the area of data sharing than it does in the area of patents and licensing. At the same time, because data sharing has been a prominent and recurrent source of tension in the global biomedical research community, CIRM has a rich history outside the state of California upon which to draw. Prior experience with data sharing in federally-funded research and multinational research efforts, such as the HGP and the HapMap Project, offers both instructive examples and cautionary tales. Achieving CIRM's multiple goals will require considerable creativity. However, if the CIRM data sharing experiment works successfully, aspects of its policy may serve as a model for other states or even for the federal government.

96. Zeeya Merali & Jim Giles, *Databases in Peril*, 435 NATURE 1010 (2005).