2

Leviathan in the Commons: Biomedical Data and the State

Jorge L. Contreras

The US federal government's 2017 budget allocated $72.4 billion to non-defense-related research and development (Holdren 2016: 3). This significant level of government-funded research generates vast amounts of data every year. In accordance with their public missions, many federal agencies have long made much of this data accessible to the public. It is widely acknowledged that the public release of data can have significant spillover effects promoting scientific discovery, technological development, and economic growth (OECD 2015: 38; Frischmann 2012).

During the Obama administration, the federal government adopted a strong commitment to the public dissemination of federally funded data in the United States (Holdren 2013; White House 2013). As explained in a 2014 White House report, "treating government data as an asset and making it available, discoverable, and usable – in a word, open – strengthens democracy, drives economic opportunity, and improves citizens' quality of life" (Exec. Off. President 2014: 11). Accordingly, the US government appears ready to foster the growth of public data resources on an unprecedented scale. But how can it do so most effectively, sustainably, and with the greatest scientific impact?

This chapter analyzes the role that the state has played with respect to the generation and management of scientific data repositories, situating it within the context of commons theory and the organization of common pool resources. Analyzing the functional roles that the state plays in data-generating research can yield a number of insights. First, the state's role in fostering innovation and scientific advancement is often analyzed in terms of incentives that the state may offer to private actors. These incentives include tax credits, intellectual property protection, grant awards, and

Jorge Contreras is Professor, University of Utah S.J. Quinney College of Law, Adjunct Professor, University of Utah School of Medicine, Department of Human Genetics; Senior Policy Fellow, Program on Intellectual Property and Information Justice, American University Washington College of Law. This chapter benefited substantially from feedback and discussion at the 2014 NYU Workshop on Medical User Innovation and Knowledge Commons. Research assistance by Michael Eixenberger is gratefully acknowledged. The author discloses that he served as legal counsel to the SNP Consortium and the International SAE Consortium, and as a member of the advisory councils of NIH's National Human Genome Research Institute (NHGRI) and National Center for Advancing Translational Science (NCATS). He currently serves on NIH's Council of Councils.

https://doi.org/10.1017/9781316544587.003 Published online by Cambridge University Press

prizes (NRC 2014: 53–56; Sarnoff 2013; Scotchmer 2004: 31–58). The functional analysis presented in this chapter reconceptualizes the state's role from that of an external actor seeking to incentivize behavior within an innovation system to that of one actor/stakeholder among many within that system.

Analyzing the functional roles of the state in the context of particular data-generating projects also highlights areas in which the state's involvement may be inefficient or ineffective in achieving its ends. As a result, suggestions may be made for improvement, both in terms of efficiency and the pursuit of specified goals. The analytical framework described in this chapter offers a means by which state engagement with data-intensive research projects may be compared across agencies, fields, and national borders. This framework can then be used to assess the effectiveness of state engagement in such research and to improve planning for future research endeavors.

2.1 COMMONS THEORY AND THE ROLE OF THE STATE

2.1.1 *Physical Resource Commons and the State*

Garrett Hardin and contemporary theorists, responding to the threat of a "tragedy of the commons," believed that the most reliable way to avoid over-consumption of finite environmental resources was through governmental intervention (Hardin 1968: 1244; Ostrom 1990: 8–12). In particular, they argued that state control and allocation of scarce resources was the only way to ensure efficient consumption and avoid environmental collapse. As explained by David Ehrenfeld, if "private interests cannot be expected to protect the public domain then external regulation by public agencies, governments, or international authorities is needed" (Ehrenfeld 1972: 322).

In contrast, theorists such as Harold Demsetz and Robert J. Smith favored an approach rooted in principles of private property. They argued that the creation of property rights in limited resources such as land would motivate property owners, acting in their own self-interest, to make the most efficient use of those resources (Demsetz 1967: 347). As Smith argues, "the only way to avoid the tragedy of the commons in natural resources and wildlife is to end the common-property system by creating a system of private property rights" (Smith 1981: 467).

Both of these approaches were challenged beginning in the 1970s by Elinor Ostrom and others, who observed numerous arrangements in which local populations made use of common resources without destroying them, and without recourse either to centralized state control or privatization (Benkler 2013: 1508; Ostrom 1990: 18–19). Ostrom's principal insight, based on these observations, was that self-governing, self-organizing systems for common property management frequently arose to address problems of scarcity, without a need for state or market-driven intervention (Ostrom 1990: 10; Rose 1986: 719–20).

Ostrom was not, however, insensitive to the participation of state actors in the management, governance, and usage of common pool resources. Under her well-known adaptation of the Institutional Analysis and Development (IAD) framework, state actors may interact with other stakeholder communities in various action arenas, and they may influence rule making with respect to common resources. For example, she describes the positive role played by "large-scale supportive institutions" such as the US Geological Survey (USGS) (Ostrom 2005: 278–79). The USGS, she explains, is uniquely positioned to offer expert technical services to local groups to assist them with the management of local resources. It would be impossible for individual groups to replicate the expertise and resources of the USGS, suggesting that the functions performed by such an agency are ideally situated within a centralized and far-reaching governmental organization.

2.1.2 *The State and Knowledge Commons*

In the mid-1990s, scholars began to apply commons theory to intangible shared resources and information (Hess and Ostrom 2007; Scotchmer 2004: 31–40). Since then, much has been written about so-called information or knowledge commons of resources such as open source software, network capacity, artistic content, academic scholarship, and scientific data (Benkler 2013: 1509, 1513–18; Madison, Frischmann, and Strandburg 2010; Hess and Ostrom 2007; Boyle 2003: 44–49). But unlike finite physical resources, information may be consumed by an unlimited number of individuals without being depleted: it is "nonrivalrous." Thus, Hardin's "tragedy of the commons," which arises from self-interested over-consumption of a finite resource, is unlikely to occur within the context of information commons. Likewise, the resource scarcity that drives theorists toward state-centric and market-centric solutions does not arise naturally in knowledge commons.[1] What role, then, can and should the state play in the creation and maintenance of knowledge commons? The next section turns to the growth of public repositories of scientific data, and the roles that the state has played in creating and maintaining these repositories.

2.1.3 *Beyond the Big Science Paradigm: Nine Roles for the State*

Since the end of World War II, the US federal government has funded scientific research projects that have generated large quantities of data. These projects, typically involving large-scale, resource-intensive, multiyear undertakings, have been made possible by government investment in advanced instruments and facilities such as particle accelerators, telescopes, and spacecraft (Scotchmer 2004: 16–26; IOM 2003: 29–79; Galison 1992). The substantial bodies of data generated by these projects have

[1] But see Frischmann (2013: 397–400), analogizing free rider problems, incentive deficits and underproduction affecting knowledge commons to the resource depleting tragedies to which physical resource commons are susceptible.

https://doi.org/10.1017/9781316544587.003 Published online by Cambridge University Press

often been made available to the public, either directly by governmental agencies (NASA's National Space Science Data Center, the National Center for Atmospheric Research (NCAR), and the US Geological Survey's Earth Resources Observation Systems (EROS) Data Center) or by government-funded repositories at private institutions (the Space Telescope Science Institute at Johns Hopkins University) (Scotchmer 2004: 240–42; NAS 2002).

The conventional "big science" account of governmental engagement in research portrays the state as either the direct generator of data or the principal funder and procurer of data from institutional researchers (Scotchmer 2004: 16–22, 228–35, 240–42; Galison 1992). Much of the policy and economics literature that considers the state's role in scientific research often focuses on this procurement function and how governmental policy can incentivize research output to maximize social welfare (Bozeman 2000: 627–55; Loiter and Norberg-Bohm 1999: 85–97).

In recent years, however, the government's role with respect to the creation and maintenance of large scientific data pools has evolved and diversified, particularly in the area of biomedical research. As a result, the traditional big science model of state research sponsorship is incomplete. As Mazzucato observes, the modern state has at its disposal a range of tools including procurement, commissioning, and regulation that it can use "to shape markets and drive technological advance" (2013: 74). This section explores the evolving role of the state in biomedical research data commons, both to improve the theoretical understanding of institutional commons structures and to inform the discussion around structuring future biomedical data commons.

First, it is necessary to identify the different roles that state actors may play in the formation and maintenance of data commons. The following nine categories offer a functional breakdown of these roles:

1 ***Creator*** The state itself, through the use of government-owned and -operated instruments, facilities, and resources, collects and generates data. This role reflects the traditional "big science" model of state-sponsored research and the many important data-centric projects historically led by national laboratories and agencies such as DARPA, NASA, and NOAA (Scotchmer 2004: 16–22; IOM 2003: 29–79).
2 ***Funder*** The state funds the collection and generation of data by academic and other non-state institutions either through grant-based funding, direct contract, or other procurement mechanisms (Scotchmer 2004: 247; IOM 2003: 82–115). The funding agency may exert varying degrees of control over the activity of the researcher and the data generated. The National Institutes of Health (NIH) is currently the largest public funder of biomedical research in the world, with an annual research budget in excess of US$30 billion (NIH 2017).
3 ***Convenor*** The state facilitates the formation of collaborative activities among private sector actors and/or governmental agencies. These "public-private partnerships" have become the focus of growing scholarly and policy attention, given their potential to harness private sector resources and expertise to solve

scientific problems prioritized by governmental agencies (Vertinsky 2015: 110; Strandburg, Frischmann, and Cui 2014). For example, the Foundation for the National Institutes of Health (FNIH), a Congressionally chartered nonprofit organization, is expressly authorized to serve as a "neutral convenor of NIH and other partners" for the purpose of encouraging dialog and collaboration between government and the private sector (FNIH 2015).

4 ***Collaborator*** The state, acting through a research-based agency such as NIH, DOE, or NASA, is an active participant in a research project that is led principally by private actors. These arrangements often arise in the context of the public-private partnerships described in the previous category but may also arise independently through interaction between researchers at government laboratories and academic institutions or private firms, or as conditions of government grants (Strandburg, Frischmann, and Cui 2014: 175–76).

5 ***Endorser*** The state is a nonparticipant that encourages a particular private sector research activity, either explicitly, by association of a state agency with the announcements and achievements of the activity, or through the implicit (and nonbinding) promise of favorable regulatory treatment for private sector participants in the activity. In this manner, the state promotes social and organizational norms (such as broad data sharing) without direct rule making or the expenditure of significant governmental resources (Rai 1999: 147–51). Through such endorsements, the state also encourages private sector behavior that is consistent with governmental goals and attitudes, with the incipient threat of greater state regulation or intervention if the private sector fails to comply.

6 ***Curator*** A state agency acts as the host and manager of data resulting from a research project and often oversees its dissemination to the public. Knowledge curation can include a range of activities including quality control, validation, preservation, collection, evaluation, and distribution (OECD 2015: 194–95; Madison 2011: 1963). An agency engaged in active curation may verify, correct, annotate, organize, and recombine data that is deposited in a repository, whereas a more passive curator may simply offer a publicly accessible location from which data may be viewed and accessed in more or less its original form (Contreras and Reichman 2015).

7 ***Regulator*** The state develops and implements policies and rules governing access to and use of a pool of scientific data (i.e., the "rules-in-use" modeled by Ostrom and others (Ostrom 1990: 50–54)). These rules may include both exogenous laws and regulations enacted by governmental authorities, as well as endogenous norms and policies imposed through the private governance mechanisms of the community.[2] Exogenous rules (e.g., laws governing intellectual property and data privacy) generally have effects beyond a single

[2] Madison, Frischmann, and Strandburg (2010: 684–88) consider the background legal regime affecting a commons to be part of the "natural" environment in which the commons exists. Their reasoning is that for intellectual resources, such as data, the legal regime often delineates the resources themselves

resource,[3] whereas endogenous rules are typically directed to the use, maintenance, and governance of the specific resource at hand.

8 ***Enforcer*** The state may police compliance with both endogenous and exogenous rules and issue formal and informal sanctions against violators. While this function is to a large degree inherent to the state apparatus, Ostrom, Rose, and others have expressed skepticism regarding central enforcement mechanisms for common pool resources (Ostrom 1990: 10; Rose 1986: 719–20). Specifically, Ostrom observed that for the state to enforce rules accurately and effectively, it must possess the capabilities to gather and evaluate large quantities of information, to monitor compliance by multiple actors, and to impose sanctions reliably and fairly. Needless to say, these criteria are not always met in practice.

9 ***Consumer*** State agencies may utilize the data found in a repository for their internal research purposes or in support of their regulatory, enforcement, and other missions. For example, as discussed in Section 2.3.2.4, the US Food and Drug Administration (FDA) utilizes data provided by applicants and clinical trial sites in assessing the safety and efficacy of drugs and medical devices submitted for regulatory approval.

Table 2.1 summarizes the nine functional roles of the state in creating and maintaining scientific data commons.

TABLE 2.1 *Roles of the state in scientific data commons*

State Role	Description
1. Creator	Generator of data through government-owned or -operated instruments
2. Funder	Funder of academic or private sector research institutions that generate data
3. Convenor	Convenor of private sector actors and governmental agencies for the purpose of engaging in collaborative research activities
4. Collaborator	Active participant in a research project involving private sector actors
5. Endorser	Nonparticipant encouraging particular private sector research activities, either explicitly or implicitly
6. Curator	Host and manager of scientific data repositories
7. Regulator	Drafter and implementer of policies and legal rules governing access to and use of scientific data
8. Enforcer	Enforcer of the above policies and rules
9. Consumer	User of data for governmental regulatory and other purposes

and their relation to the contextual environment. This differs from the natural resource context examined by Ostrom.

[3] Depending on the size and nature of a common pool resource, some exogenous rules may target it exclusively. For example, governmental permitting requirements, land use regulations, zoning ordinances, and the like may be narrowly tailored to affect specific properties or resources to the exclusion of others (Ostrom 1990: 50–54).

The existence of these overlapping and complementary state roles, while underappreciated in the literature of scientific research, is not surprising when data and public repositories of scientific data in particular are viewed as elements of the scientific research infrastructure. Generally speaking, infrastructural resources such as roads, communication systems, and utilities provide the "underlying foundation or basic framework" needed to support a wide range of downstream productive activities (Frischmann 2012; NRC 1987). As such, scientific data repositories that are broadly accessible may be considered key elements of the scientific research infrastructure (OECD 2015: 179).

The state plays a number of well-understood roles with respect to the planning, provisioning, and maintenance of publicly owned infrastructure resources such as highways, prisons, and public utilities. Likewise, the state is often involved in the oversight, regulation, and operation of private and public-private infrastructural resources such as airports and telecommunications networks. Why then should the same types of complementary and overlapping relationships not arise with respect to data resources that form an integral part of the research infrastructure?

In the case studies that follow, the evolution of the state's role from big science provisioner to the more multifaceted relationships described earlier is analyzed.

2.2 NIH AND THE GENOME COMMONS

The multiple roles of the state described in the preceding section are illustrated by the US federal government's involvement in the creation, growth, and ongoing maintenance of the large body of public data concerning the human genome. This aggregation of data, which has been referred to as the "genome commons" (Contreras 2011, 2014) presents a useful case study for several reasons. First, the genome commons, which was initiated in the late 1990s with the Human Genome Project (HGP), has had a long and well-documented history (McElheny 2010; Contreras 2011). Over the two decades of its existence, it has adapted to accommodate a range of organizational and institutional changes, within both the government and the larger biomedical research community. The genome commons, which today includes data from a broad range of public, private, and public-private research efforts, can also be characterized as a common pool resource of the type described by commons theorists. That is, the genomic and associated data contained within the commons is provisioned and governed through a set of polycentric, multi-stakeholder mechanisms (Contreras 2014: 107–08).[4]

[4] The focus of this chapter is on biomedical data resources. For the sake of brevity, it does not seek to address the constellation of related issues surrounding physical biological samples held in hospitals, laboratories, biobanks, and repositories around the world. For a comprehensive discussion of these issues, see, e.g., Reichman, Uhlir, and Dedeurwaerdere (2016) and Rhodes (2013).

2.2.1 *Beginnings: The Human Genome Project*

Though researchers began to identify bits and pieces of the human genetic code in the mid-twentieth century, the creation of a complete sequence map of the human genome was not seriously proposed until 1985. At that time, leading genetic researchers, encouraged by the emergence of improved DNA sequencing technologies, first outlined a proposal to sequence the entire 3.2 billion DNA base pairs constituting the human genome. Doing this, they argued, would lead to significant improvements in understanding the genetic bases of disease (McElheny 2010: 17–33; Cook-Deegan 1994: 79–91). These efforts opened the door for the largest biomedical research endeavor of its day, the HGP.

Two US federal agencies initially led the HGP: the NIH, the funder of most disease-focused research in the United States, and the Department of Energy (DOE), whose expertise in genetics arose from studying the effects of radiation on atom bomb survivors (Cook-Deegan 1994: 97–104). These two agencies joined forces in 1990 to co-lead the project with additional support from the UK-based Wellcome Trust and funding agencies in the United Kingdom, France, Germany, and Japan. The massive research effort was compared to the Manhattan Project and the Apollo lunar landing program, among other projects. Yet even at an early stage, the role of the state in the HGP was more complex and multifaceted than it had been in these previous "big science" endeavors.

State as Convenor Unlike earlier large-scale scientific projects relating to national defense and space exploration, the proposal to map the human genome originated with academic researchers rather than government officials. From an early stage, leaders at NIH and DOE interacted closely with the scientists who proposed the project and who would eventually carry it out. As the HGP coalesced, governmental actors worked closely with academic investigators not only to develop a scientific roadmap for the project but also to establish rules regarding the sharing and release of data generated by the project (see later discussion of "State as Regulator").

Moreover, rather than assign career bureaucrats to oversee the HGP, NIH recruited prominent scientists to lead the project. Chief among these was James Watson, Nobel laureate and co-discoverer of the structure of the DNA molecule, who was appointed to oversee the newly formed National Center for Human Genome Research. Watson was succeeded in 1992 by Francis Collins, a prominent genetic researcher from the University of Michigan who was best known for his leading role in discovering a gene closely associated with cystic fibrosis. A host of other prominent researchers, in the United States, the UK, and elsewhere, were active in the leadership of the HGP, facilitating the close interaction of government and academia during the long project (Contreras 2011: 76–77; n.60). In this sense, NIH acted as a *convenor* of the scientific community, bringing it together to collaborate on planning and executing the most ambitious scientific undertaking of its day. Without the engagement of the broader scientific community, the HGP would never have been possible.

State as Funder/Creator The HGP is estimated to have cost more than $2 billion to complete, the bulk of which the NIH funded. NIH's research funding is typically allocated through grant awards. These awards are based on the agency's evaluation of competitive research proposals submitted by academic and other investigators.[5] Award recipients are expected to complete the projects that they have proposed, but the agency seldom intervenes in the conduct of the research program itself.

The HGP was organized differently. Rather than act as a passive research funder, NIH led planning efforts and solicited bids from research institutions to carry out specific components of the project (IOM 2003: 31–40). Three academic centers were selected to perform DNA sequencing: Washington University in St. Louis, the Whitehead Institute at MIT, and the Sanger Centre in Cambridge, England. Thus, even though NIH did not carry out the sequencing work using government-owned or -operated resources, it assumed a leading role with respect to the data generated by the HGP that could characterize it as a *creator* as much as a *funder*.

State as Curator Although the sequencing work of the HGP was carried out by academic research institutions funded by NIH and the other project sponsors, hosting and maintenance (curation) of the massive (for the time) quantities of data generated by the project fell to NIH itself.[6] This curatorial role was not unique to the HGP. Over the years, governmental projects in fields such as astronomy, earth science, and particle physics made large quantities of observational and experimental data available to the public. This data was often hosted at federally managed facilities such as the National Center for Atmospheric Research (NCAR) and the US. Geological Survey's Earth Resources Observation Systems (EROS) Data Center, and at private institutions contracted by the federal government, such as the Space Telescope Science Institute at Johns Hopkins University (NAS 2002). The HGP was distinctive, however, in that the data being curated by the state was not generated by government-owned and -operated instruments, but by academic institutions supported by governmental funding. Thus, in the case of HGP data, the state's role as curator diverges from the role it typically assumed in big science projects.

The HGP elected to utilize the existing GenBank database, administered by the National Center for Biotechnology Information (NCBI), a division of the NIH's National Library of Medicine, for the deposit and public release of genomic sequence data. GenBank originated with the Los Alamos Sequence Library operated by Los Alamos National Laboratory since 1979 (Hilgartner 1995: 243). NIH contracted with Los Alamos in 1982 to create GenBank as a publicly accessible repository for DNA sequences (Strasser 2008: 538). It has been operated by NCBI

[5] An example of such grant-funding mechanisms is discussed in Strandburg, Frischmann, and Cui (2014).

[6] The European Molecular Biology Library (EMBL) and DNA Data Bank of Japan maintain comparable repositories that are synchronized with NCBI's GenBank on a daily basis (Strasser 2008).

since 1992, making it one of the longest-running state-operated repositories of scientific data (Ankeny and Leonelli 2015: 128–29; Benson, Lipman, and Ostell 1993: 2963).

In addition, given its pre-HGP origins, GenBank is, and has always been, open to deposits of DNA sequence information from sources other than NIH-funded research projects. Underscoring this source-neutral policy, GenBank even accepted the human genome sequence data generated by privately held Celera Genomics, which competed fiercely with the HGP to sequence the human genome from 1998 to 2000 (Kaiser 2005: 775).

State as Regulator From an early date, NIH's involvement in the HGP included the creation of both exogenous legal regulations and endogenous rules-in-use. One of the most important sets of exogenous rules affecting the HGP concerned the ability of private parties to obtain patents covering DNA sequence information. By the mid-1990s, many commentators feared that allowing patents on DNA sequences of unknown function would stymie biomedical research (Heller and Eisenberg 1998: 698; Cook-Deegan 1994: 308–11). This perceived threat was one of the most hotly contested legal issues in the emerging field of genomics and "became the main focus of a cottage industry of biotechnology patenting articles in law reviews and scientific journals" (Demaine and Fellmeth 2002: 326).

NIH's position, which it solidified only after a contentious attempt to obtain its own patents covering DNA sequences, strongly disfavored the patenting of genetic material. Beginning in the mid-1990s, the agency engaged in an unofficial campaign to persuade the US Patent and Trademark Office (USPTO) to stop issuing such patents (NRC 2006: 52–53). Based on this and other input, in 1999 the USPTO adopted a policy that disallowed the patenting of DNA sequences of unknown function because they lack the required "utility" for patent protection (USPTO 1999: 714–40).

Perhaps even more significantly, NIH led the development of the endogenous rules-in-use that governed the deposit and release of genomic data generated by the HGP. NIH's 1996 data release policy for the HGP was largely based on the so-called Bermuda Principles, a set of guidelines developed by a group of prominent researchers and policymakers (Bermuda Principles 1996). The Bermuda Principles were revolutionary in that they established, for the first time, that data from public genomic projects should be released to the public almost immediately after being generated, rather than after a waiting period of 6 to 18 months, as was the norm for federal projects at the time (Contreras 2011: 84–85; Bermuda Principles 1996; NHGRI 1996). These rapid data release requirements were intended to promote three NIH policy goals: achieving coordination among the many independent sequencing centers working on the HGP, accelerating scientific advancement, and limiting third parties' ability to patent data first generated by the HGP (Contreras 2011: 86).

2.2.2 *The State's Evolving Role in Post-HGP Genomics Projects*

The HGP published its first draft of the full human genomic sequence in 2001 and announced its completion in 2003. As the lengthy public project drew to a close, researchers began to plan a number of follow-on activities designed to build on and make use of the basic sequence data generated by the HGP. These projects included both NIH grant-funded projects along the lines of the original HGP (these projects included the Encyclopedia of DNA Elements (ENCODE) (2003), the Cancer Genome Atlas (TCGA) (2006), and the Human Microbiome Project (HMP) (2007)), as well as projects in which NIH partnered with a range of public and private sector funders, both from the United States and abroad (Contreras 2011: 97–107, 2014: 123–27). Particularly in this later category of projects, NIH's role shifted from that of a typical big science *creator* and *funder* to that of a *convenor*.

Expansion of the Convenor Role Shortly after the HGP released its initial draft sequence in 2001, a group of researchers led by Eric Lander at the Whitehead Institute proposed a project that would chart the ways in which markers along the human genome recurred in groups (haplotypes) (Int'l HapMap Consortium 2003: 789). Though NIH participated in funding the resulting "HapMap" project, additional funding came from state agencies in Japan, the United Kingdom, Canada, China, and Nigeria. The HapMap project was in no sense "led" by NIH or the US government. Rather, NIH served, to a degree, as a convenor of other parties, both governmental and academic, that were interested in the project.

Curation as an Independent Function The GenBank database maintained by NCBI served as one of the principal repositories for DNA data generated by the HGP. NCBI's curatorial role with respect to this data was similar to that played by other governmental agencies that maintained large data sets that they created or funded. At a high level, the genomic data managed by NCBI was not so different from radio telescope data managed by NASA or atmospheric data managed by NOAA.

But following the conclusion of the HGP, as the cost of gene sequencing began to decrease, more and more human and nonhuman genomic data was produced by researchers around the world. Though many of these researchers were unaffiliated with NIH's large-scale data-generation projects, they, too, were welcome to deposit genomic data in GenBank at no charge. GenBank thus became a global repository for DNA sequences and related data, irrespective of their origin, and NCBI became the de facto curator of this data from sources around the world.

Far from being a passive role, curation of genomic data requires not only suitable computing and networking resources but also significant technical and scientific expertise in data selection, quality control, formatting, display, and visualization (Ankeny and Leonelli 2015: 133–34; OECD 2015: 194–95; Madison 2011: 1982–87). Academic research groups curate many important scientific

databases, and governmental agencies stepping into this role have comparable requirements. For example, the DNA sequence data uploaded by researchers to GenBank may at times be duplicative, incomplete, or corrupted. Researchers wishing to download the complete genome of a particular organism would be hard pressed to identify and assemble all the necessary elements from GenBank deposits. NCBI addressed this issue with the introduction of the RefSeq (reference sequence) database in 2000. RefSeq contains a "reference" genome for each organism (and particular strains of organisms) that is compiled by NCBI staff from GenBank records. RefSeq genomes are continually updated and refined as higher-quality data is added to GenBank (Lee, Chapter 3, this volume; NCBI 2013).

Another significant aspect of curation in the area of biomedical data involves the protection of individual health information. Like astronomical and atmospheric data, it was once thought that DNA sequence data divorced from its donors' identities (de-identified) carried little risk to individuals. But as genomic research evolved and researchers began to explore the associations between genes and health, they began to link genomic sequence data with physiological, demographic, and clinical data (phenotypic data). While genome-wide association studies (GWAS) have shed substantial light on the interactions between genes and human health, they also give rise to increased risks that the identities of DNA donors can be determined from disclosed data. Today, in fact, many researchers believe that it is virtually impossible to de-identify genetic data with absolute assurance (e.g., Couzin-Frankel 2015: 502).

To address the risks of re-identification of data subjects and to accommodate the linkage of DNA sequence data with phenotypic data, NCBI created the Database of Genotypes and Phenotypes (dbGaP). NCBI's curatorial role with respect to dbGaP is significantly greater than it is with respect to GenBank: dbGaP has a two-tiered structure that allows access to potentially identifying information to be authorized on a case-by-case basis by a standing Data Access Committee (DAC) composed of NIH personnel (Paltoo et al. 2014: 936). Through this approval function, NIH acts not only as a technological curator of data but also as the guardian of personally sensitive information that may be gleaned from the data stored within its repositories.

While the state, through the curatorial role played by NCBI and other NIH divisions, provides an invaluable service to the global scientific community, this service is not without cost. NCBI, which operates GenBank and a number of more specialized data resources, had an annual budget in 2016 of approximately $54.3 million (NLM 2016). Other NIH divisions support a wide range of biomedical research databases. It has been estimated that the annual budget for NIH data resources excluding NCBI and other NLM divisions is approximately $110 million (Kaiser 2016). At these funding levels, NIH has been under pressure to reduce its support for more specialized data resources, including genomic repositories for various microorganisms and model organism systems (Kaiser 2016). Thus, even in the area of genomics, where the state

has been a leader in developing and curating valuable data resources, the pendulum may be swinging back toward a more modest role for state actors, both in terms of fewer supported data resources and a lower level of curation and maintenance for those data resources that remain (Kaiser 2016; Mishra, Schofield, and Bubela 2016: 284; Contreras and Reichman 2015: 1312).

Increasing Regulation Following the HGP, both the expanding types of data housed within genomic data repositories and the privacy risks associated with individual health data led to increasingly detailed and complex endogenous rules governing genomic data resources. NIH's policies grew from relatively simple Bermuda-based requirements regarding the timing of data deposits to comprehensive regulations, exemplified by NIH's 2007 GWAS policy, regarding data security, access, and usage, as well as the ability of investigators to publish discoveries made using genomic data and to seek patents claiming those discoveries (Contreras 2014: 123–29).

The growth in policy length and complexity, however, does not necessarily indicate a shift in NIH's role from collaborator to regulator. As I have described previously, the development of NIH's policies regarding genomic data resulted from an open and consultative process among multiple stakeholders including researchers, patient advocacy groups, and private industry (Contreras 2014: 107–11, 127). Moreover, as noted earlier, many of the NIH officials involved in policymaking are themselves respected scientific researchers with significant and ongoing research programs. Thus, NIH's increasing codification of the rules-in-use of the genome commons does not necessarily detract from its role as collaborator. The same may not be true, however, with respect to the agency's role as "enforcer," discussed next.

State as Enforcer As the developer and implementer of rules governing the deposit, access, and use of genomic data housed in NIH repositories, NIH stands in a unique position to monitor and enforce compliance with those rules. Thus, if an NIH-funded sequencing center failed to upload its data to GenBank within the required time period, NIH could take a number of enforcement steps including discussing the deficiency with the delinquent center; developing a remedial plan; and if repeated violations occurred, withholding or reducing the funding to that center.

Many NIH rules-in-use, however, are phrased as "encouragements" rather than "requirements" (sometimes referred to as "norms" or "soft" rules) (Contreras 2011: 87–88; Ostrom 2005: 121–27; Rai and Eisenberg 2003: 293–94). NIH's policy discouraging the patenting of DNA sequences is one example of such "soft" rules. This policy warns that the agency "will monitor grantee activity ... to learn whether or not attempts are being made to patent large blocks of primary human genomic DNA sequence" (NIH 1996). With respect to its dbGaP database, NIH catalogs the types and frequency of policy violations that it discovers, including data submission errors, inappropriate use or dissemination of data, data security lapses, and violations of publication embargoes (NIH 2015).

Yet it is not clear how effective NIH's policing function has been, even with respect to dbGaP. The agency claims that with more than 20,000 data access requests between 2007 and 2015, it has identified only 27 policy violations, all of which it has managed to a satisfactory resolution (NIH 2015). Several of the 27 incidents were reported to NIH by the violators themselves; others were caused by bugs in the dbGaP software and procedural errors. This low rate (less than 0.1% of total data access requests) could indicate either a low incidence of noncompliance or, more likely, a low incidence of *detection*. The handful of disclosed noncompliance incidents offer little indication that the agency has implemented an effective program to monitor and police the use of genomic data.

Moreover, unlike other federal enforcement agencies that make their investigations and conclusions public, NIH, despite its rhetoric of openness, does not publicly disclose the names of parties or individuals implicated in its reported policy violations. This hesitancy may offer a clue as to the underlying causes of NIH's weak enforcement of its genomic data policies. Ironically, it is the very diversity of roles played by the agency in the research enterprise that may hamper its desire or ability to enforce its rules vigorously. That is, unlike a neutral state enforcement agency such as the Department of Justice, NIH has numerous institutional ties to its funded researchers: NIH officials are drawn from NIH-funded research institutions; NIH has convened many of the relevant research groups, tying its reputation to the success of the project; NIH staff (intramural researchers) collaborate closely with extramural NIH-funded researchers; and NIH eventually holds and curates the data produced by the research effort. These close ties may make NIH officials reluctant to enforce the agency's rules against familiar research institutions and researchers, leading perhaps to more informal (and possibly less effective) enforcement after the occurrence of actual and suspected violations.

2.2.3 *Public-Private Genomics*

In addition to the federally funded genomics projects described above, a number of significant private sector genomic research projects emerged during and after the HGP. However, unlike typical industrial research programs, many of these initiatives publicly released large quantities of genomic data, to both governmental and privately operated repositories.

Convenor and Collaborator Even though the federal government is not the primary funder or planner of private sector research initiatives, it may engage with them in several ways. NIH in particular has an active program of collaborating with the private sector through public-private partnerships, and through its National Center for Advancing Translational Sciences (NCATS) (NCATS 2015; NIH 2010). The US Food and Drug Administration (FDA) also encourages and participates in collaborations with the private sector through its Critical Path Initiative, among other programs (FDA 2017).

One example of a successful public-private collaboration in the area of genomics data was the Genetic Association Information Network (GAIN). GAIN was launched in 2006 as a public-private partnership among commercial firms (Pfizer, Affymetrix, Perlegen Sciences, and Abbott), academic institutions (the Broad Institute), NCBI, and FNIH (GAIN 2007: 1045–46). The goal of the project was to use genome-wide association studies (GWAS) to study the genetic basis for six common diseases. GAIN involved both FNIH's role as a *convenor* and NIH/NCBI itself as a *collaborator*, a combination of roles that is not uncommon.

Endorser The SNP Consortium Ltd. was formed in 1999 by a group of pharmaceutical and information technology firms with additional financial support from the Wellcome Trust (Holden 2002: 22–26). The consortium's goal was to identify and map genetic markers known as "single nucleotide polymorphisms" (SNPs) during the concluding years of the HGP. Though NIH did not formally join or fund the SNP Consortium, it actively monitored its activities and helped coordinate the consortium's research with the data being produced by the HGP. Given the HGP's highly publicized race with Celera Genomics, NIH welcomed and publicly supported private sector research activities that worked in concert with, and did not oppose, the public HGP. In this sense, NIH acted as a significant *endorser* of the SNP Consortium and similar efforts (McElheny 2010: 143; Shreeve 2004: 294).

The state's endorser role was more recently exemplified by the FDA's interaction with the International Serious Adverse Events Consortium (iSAEC), a group of pharmaceutical and health care companies organized in 2007 to identify DNA markers associated with serious adverse drug reactions (Holden et al. 2014: 795). The FDA helped generate industry support for the iSAEC and its mission, which is in line with recent FDA initiatives relating to drug safety (Holden et al. 2014: 795). Though no formal relationship exists between the FDA and iSAEC, the agency sends representatives to iSAEC meetings and jointly announces research milestones with iSAEC (US Food and Drug Admin. 2010). As a result, iSAEC's activities are portrayed to the public as aligned with the FDA, thereby validating the organization and its activities. The FDA likewise benefits from association with a research program that has generated significant data in a field that is important to the agency's mission.

Curator In addition to interactions with the state in the conduct of research programs, private sector researchers often submit genomic data to federally supported databases such as GenBank and dbGaP. NCBI will accept and curate this data at no charge to the submitter while offering substantial technical expertise and oversight. As a result, NCBI plays a significant *curatorial* role with respect to privately sourced data. As noted earlier, even Celera Genomics eventually deposited its entire human and mouse genome sequences in GenBank (Kaiser 2005: 775). The SNP Consortium, which made its data available through a privately operated website, also uploaded this data to GenBank, as did a significant genomic research effort sponsored by the pharmaceutical firm Merck (Contreras 2011: 95; Holden 2002: 25–26). GAIN, along

with other public-private research collaborations, also deposited its data in dbGaP (GAIN 2007).

Thus, while many private sector researchers retain their data within corporate repositories, those that release data to the public often do so through state-supported facilities such as GenBank and dbGaP. It is likely that this route is attractive to private sector researchers, as NCBI conducts its massive data curation program at taxpayer expense, while providing substantial bioinformatics and data curation expertise. It is in the state's interest to offer this service to maximize the likelihood that data will be utilized by a broad range of researchers, thus advancing scientific progress, and also ensuring that the data generated by private sector researchers will be subject to the same data quality, privacy, and security restrictions as data generated by state-funded projects. As a result, a symbiotic relationship is fostered between private sector and state actors in the area of data curation.

Nevertheless, as noted earlier, the cost of maintaining an ever-expanding set of data resources has already begun to strain federal budgets. As a result, the future may see a shift back toward privately curated data collections in some areas. The hope, from a social welfare standpoint, is that the custodians of these private resources will continue to make them broadly accessible to the public, even with reduced state support.

Limited State Regulation and Enforcement Though exogenous laws and regulations impact private sector research to much the same degree as they impact state-funded research, NIH did not play a direct role in formulating the endogenous rules-in-use of private sector genomic data projects. Rather, these rules have typically been created by the institutions and firms involved in a project, based on their internal goals, requirements, and policies. Nevertheless, the influence that NIH rules regarding data access and use have on private sector policies is significant. First, NIH's data policies have become norms in the field of genomics, if not the broader biomedical arena. Researchers in the private sector have often been educated and trained at academic institutions and have thus become accustomed to the requirements of such policies, internalizing these norms in their standard scientific practices. Moreover, most academic institutions are recipients of federal research funding, and many have officially or unofficially adopted internal rules and policies for data sharing that conform to federal standards. As such, the state has acted as a norm setter, even in these private sector research projects.

2.2.4 *Assessing the State's Roles in Genomics Research Projects*

Table 2.2 illustrates the roles played by the state in genomic research projects that have resulted in the contribution of data to the public, including the HGP, post-HGP federally funded genomic research projects, and private sector projects that have contributed genomic data to the public.

TABLE 2.2 *Roles of the state (NIH) in genomic data generation projects*

State Role	HGP	Post-HGP (NIH)	Private Sector
Convenor	Yes	Sometimes	Sometimes
Funder	Yes	Yes	No
Creator	Partially	Partially	No
Collaborator	Yes	Sometimes	Sometimes
Endorser	Yes	Yes	Sometimes
Curator	Yes	Yes, but declining	Sometimes
Regulator	Yes	Yes	Yes, if data is state curated
Enforcer	Yes (though never exercised)	Yes (though rarely exercised)	Yes, if data is state curated
Consumer	No	No	No

As Table 2.2 illustrates, NIH has played a range of roles in these projects, going well beyond that of either passive funder or big science creator. These additional state roles have added substantial value to data-generating research projects. In some cases, research may not have been possible, or would have had a less significant impact, were it not for the supporting roles played by the state. Accordingly, to maximize the effectiveness of a government-funded data-generating project, planners should take into account the different potential roles of state actors over the life cycle of the data.

Even more interesting are the many potential roles that the state may play in public-private or private sector data-generating projects. The genome commons provides several examples in which substantial benefits have accrued from the state's convening of multiple independent actors, its collaboration with academic and industrial researchers, and its curation of large data sets in existing or new data repositories.

Despite these benefits, as discussed in Section 2.2.2, NIH's enforcement of its data access and usage rules has been weak, possibly as a result of a failure to detect violations or a failure to initiate enforcement measures against violators. Thus, to the extent that either governmental or private sector planners wish to implement a robust set of rules-in-use relating to their data commons, they could explore policing and enforcement options beyond those offered by NIH's current model.

2.3 SHAPING THE ROLE OF THE STATE IN FUTURE DATA COMMONS: CLINICAL TRIALS DATA

As illustrated by the genome commons, the state may play a range of roles in the generation and maintenance of scientific data commons. These roles extend well beyond the traditional big science model of state-sponsored resource creation and provisioning. However, there is little detailed analysis of state roles at the

https://doi.org/10.1017/9781316544587.003 Published online by Cambridge University Press

outset of commons formation. This section suggests ways in which the analytical framework developed here may be used by policymakers and project planners to model the engagement of state agencies in new public aggregations of scientific data.

2.3.1 *Sharing Clinical Trials Data*

Clinical trials are health interventions designed to assess the safety or efficacy of a drug or medical device. Clinical trials are required by the FDA to obtain regulatory approval to market a new drug or medical device in the United States (IOM 2015: 68–69).[7] Such trials are often sponsored by a pharmaceutical or medical device company, which contracts with one or more academic research institutions to conduct the trials. More than 200,000 clinical trials have been conducted worldwide since 2000 (ClinicalTrials.gov 2015).

Many clinical trials involve thousands of individuals observed over long periods of time. In addition to data regarding the intervention being studied (e.g., the chemical composition and other characteristics of the drug, device or procedure, the experimental protocol, data analysis methods, and analytic code), data is collected regarding the study participants' health history, demographic profile, phenotypic traits, clinical diagnosis, adverse reactions, and post-intervention prognosis (IOM 2015: 97–105). Much of this data is submitted to the FDA in support of new drug or device applications. The Food and Drug Administration Amendments Act of 2007 (FDAAA) requires that summary data be disclosed to the public via the NIH-operated ClinicalTrials.gov website, but this data is limited to major outcomes and adverse events. Moreover, data from trials that were deemed unsuccessful and did not result in an approved drug or device are typically not required to be disclosed. As a result, the vast majority of clinical trials data that is generated remains nonpublic (IOM 2015: 113).

In view of this situation, many have argued that more clinical trials data should be made available to the public. Between 2012 and 2015, the Institute of Medicine (IOM) conducted a series of workshops that explored issues relating to the sharing of clinical trials data. It summarized the substantial public benefits of sharing this data as follows:

> From the perspective of society as a whole, sharing of data from clinical trials could provide a more comprehensive picture of the benefits and risks of an intervention and allow health care professionals and patients to make more informed decisions about clinical care. Moreover, sharing clinical trial data could potentially lead to enhanced efficiency and safety of the clinical research process by, for example, reducing unnecessary duplication of effort and the costs of future studies, reducing

[7] The focus of this section is US law and regulation. However, similar regulatory regimes exist in most developed countries. According to the OECD (2015: 339), at least 10 countries have planned to implement systems for systematically analyzing clinical trials data for public health and other purposes.

> exposure of participants in future trials to avoidable harms identified through the data sharing, and providing a deeper knowledge base for regulatory decisions.
>
> In the long run, sharing clinical trial data could potentially improve public health and patient outcomes, reduce the incidence of adverse effects from therapies, and decrease expenditures for medical interventions that are ineffective or less effective than alternatives. In addition, data sharing could open up opportunities for exploratory research that might lead to new hypotheses about the mechanisms of disease, more effective therapies, or alternative uses of existing or abandoned therapies that could then be tested in additional research. (IOM 2015: 32)

Offsetting these benefits, of course, are risks that may arise from data sharing, including compromising the privacy and confidentiality of trial participants, inadvertently disclosing sponsor companies' trade secrets, fueling spurious liability suits, and deterring the use of potentially beneficial therapies (IOM 2015: 33–34). As a result, the IOM has recommended a balancing of interests, with a goal of maximizing the benefits of sharing clinical trial data while minimizing its potential harms (IOM 2015: 34). Some private firms, together with academic institutions, have already taken steps toward broader data sharing, but these early efforts remain tentative (IOM 2015: 20–21).

Were more clinical trial data to be publicly disclosed, as envisioned by the IOM and others, this data would contribute to a substantial knowledge commons. Like the genome commons, a clinical trials data commons would most likely be governed through polycentric mechanisms involving stakeholder groups including study participants, funders and sponsors, advocacy groups, regulatory agencies, researchers, research institutions, scientific journals, and professional societies (IOM 2015: 4–6). Existing rules-in-use would need to be expanded to establish the scope of public access to such data, and the terms on which it could be utilized (IOM 2015: chap. 5).

2.3.2 *Potential Roles of the State in a Clinical Trials Data Commons*

2.3.2.1 State Actors and Primary Roles: NIH as Funder, FDA as Regulator

NIH and FDA are the two primary federal agencies involved in the conduct of clinical trials in the United States, although their roles differ significantly. NIH financially supports numerous clinical trials.[8] As of 2015, this support covered more than 3,000 active clinical trials in the United States (IOM 2015: 59). Unlike the HGP and the broader genome commons, however, NIH plays little role in the planning and conduct of clinical trials or the generation of clinical trials data. As such, its role with respect to clinical trials data is closer to that of a typical grant-based funding agency than the funder-creator role that it assumed in genomics projects.

[8] In addition to receiving support from NIH, clinical trials are supported by private industry, charities such as the Wellcome Trust and the Bill and Melinda Gates Foundation, as well as disease advocacy groups (IOM 2015: 58).

The FDA, on the other hand, serves as the principal US regulator of new drugs and medical devices. The FDA requires that applicants for approval of these new interventions conduct clinical trials to establish their safety and efficacy. As noted earlier, the FDA also requires that summary data generated by clinical trials be released to the public and is engaged in an ongoing public discussion regarding the potential expansion of the scope of data to be publicly released (IOM 2015: 173–76; Hudson and Collins 2014: 365).

2.3.2.2 Clinical Trials Data Curation

The principal public repository of summary clinical trials data today is housed at NIH's National Library of Medicine (NLM) and is accessible through the ClinicalTrials.gov website.[9] Thus, as it does with genomic data, NIH serves a curatorial function for clinical trials data. However, unlike the data stored in repositories such as GenBank and dbGaP, clinical trials data is generated by researchers without direct NIH involvement. Because clinical trial data is central to the FDA regulatory approval process and forms the basis on which FDA evaluates applicants' new drugs and devices, NIH does not validate, annotate, or enhance clinical trial data, as it does genomic data (e.g., in creating the RefSeq database). Moreover, because the types of data that are made accessible through ClinicalTrials.gov are mandated by statute, NIH does not maintain a data access committee (DAC) to review and approve applications to access clinical trial data, as it does with genomic data stored in dbGaP. Thus, while NIH performs a curatorial function with respect to both clinical trial and genomic data, its role with respect to clinical trial data is more passive and mechanical than it is with respect to genomic data.

Moreover, ClinialTrials.gov today houses only the summary data required by the FDAAA. If more substantial clinical trials data were to be released in a systematic manner, some observers fear that NCBI's current platform is inadequate and far greater computing resources and infrastructure will be required (IOM 2015: 15). As a result, substantial investments will need to be made in data infrastructure, and the role of the state in this infrastructure development, as well as ongoing data curation, will need to be clarified.

Given these considerations, planners may wish to consider alternatives to state curation of clinical trials data beyond ClinicalTrials.gov. For example, outsourcing this function to a private sector contractor may result in a lower-cost and more

[9] The statutory mandate for ClinicalTrials.gov is found in Section 113 of the Food and Drug Administration Modernization Act of 1997, which required NIH to create a public resource including information about federally and privately funded clinical trials conducted under investigational new drug applications (INDs) being prepared for submission to the FDA. Congress enhanced the requirements for submission to ClinicalTrials.gov in Section 801 of the Food and Drug Administration Amendments Act of 2007. The amendment expanded the number and types of trials requiring registration, as well as the amount and type of summary data to be disclosed.

streamlined system that does not otherwise drain resources from the budget-constrained NIH/NCBI. While such an approach could have drawbacks in the highly technical realm of genomics data, it may be a preferable solution in the case of a data system that requires only hosting and open access. Such privatized data hosting systems have been successfully implemented by agencies such as the Securities and Exchange Commission[10] and may be worth consideration as larger bodies of clinical trials data become available for public release.

2.3.2.3 Enforcement of Data Policies

Rules regarding the deposit, access, and use of data within an expanded clinical trials database will likely be promulgated by the FDA and/or Congress. As noted earlier, the FDAAA today requires that clinical trial sponsors submit summary data to ClinicalTrials.gov and imposes stiff penalties for noncompliance (fines for reporting violations can reach $10,000 per day) (ClinicalTrials.gov 2012). Nevertheless, compliance with these requirements has been referred to as "disappointing" (Hudson and Collins 2014: 355). According to IOM, only 46 percent of NIH-funded clinical trials publish their results within 30 months of completion (IOM 2015: 59). One study found that even after four years, the results from 30 percent of a sample of 400 clinical trials had neither been published nor reported on ClinicalTrials.gov (Saito and Gill 2014).

These findings suggest that current enforcement of the rules governing clinical trials data is not being managed effectively. There are several possible reasons for the enforcement gaps in this area. First, the agencies responsible for policing compliance with the rules, NIH and FDA, may lack the resources to undertake necessary compliance-monitoring measures. While a lack of resources is a perennial issue with governmental agencies, more systemic issues likely hinder effective enforcement of data-related rules. NIH, in particular, acts as a *collaborator* with the institutions conducting clinical trials, which, as discussed in Section 2.2.2, may make agency personnel reluctant to enforce rules too harshly against them. This risk is particularly acute if the same agency personnel are responsible for both collaborative activity and compliance monitoring and enforcement (i.e., both because of the greater potential sympathy that agency personnel may have for their colleagues and because agency personnel who are actively collaborating in the conduct of trials may have less need for access to data through public means, making its unavailability less noticeable and inconvenient to them). Accordingly, if the state wishes to mandate the disclosure of expanded clinical trials data, it will need to develop more robust approaches to enforcing its disclosure requirements.

[10] In 1997, the US Securities and Exchange Commission (SEC) privatized its Electronic Data Gathering, Analysis, and Retrieval (EDGAR) Public Dissemination Service (PDS) (US Securities and Exchange Comm. 1997).

2.3.2.4 State as Consumer

One distinct role that the state plays with respect to clinical trials data, but which it does not significantly play in the genome commons, is that of a *consumer* or user of the data for its internal purposes. The FDA in particular utilizes clinical trials data submitted by researchers in its regulatory capacity, as it evaluates applications for new drug and device approvals. The character of this role is distinct from other roles played by the state as it places the state in a position similar to other private and public sector users of data within the commons. Yet there are also differences between the state's use of clinical trials data and, say, an academic or industrial researcher's use of genomic data. Whereas the latter researchers typically access and use data to advance their own research, the FDA's utilization of clinical trial data supports a regulatory function that ultimately inures to the benefit of the data submitter (if the drug or device application is approved), and to society more broadly (Abbott, Chapter 6, this volume).

But while the new drug or device applicant benefits from the FDA's review of its data (assuming that review is favorable), the applicant does *not* benefit from the disclosure of its data to the public (which includes not only interested citizens but also the applicant's competitors and potential litigants over safety and other claims). The FDA, however, reviews all the applicant's trials data, whether or not the data is made public. As discussed in Section 2.3.1, the question for policymakers is how much of the applicant's data should be disclosed to the public, and the degree to which social welfare gains from disclosure outweigh potential prejudice to the applicant and privacy risks to individual trial participants (IOM 2015: 32–34). As a consumer of the data, the agency itself is relatively unaffected by the amount of data publicly disclosed.[11]

2.3.3 *Assessing State Roles: Genomics and Clinical Trials*

Table 2.3 offers a comparison of the roles of the state in the genome commons and its potential roles in a clinical trials data commons. Consideration of the different roles played by the state in these two contexts suggests ways that state involvement may be configured in new data commons that may enhance the efficiency and effectiveness of data-sharing arrangements and improve overall social welfare.

As illustrated, NIH plays a lead or strong collaborative role in many genomics data generation projects. The agency's role is less active with respect to clinical trials, however, tending more toward that of an external funder. There may, however, be opportunities for NIH to use its substantial internal expertise in support of clinical trials and data generation. Such opportunities may arise, for example, through the work of the National Center for Advancing Translational Sciences (NCATS), which

[11] One could even argue that the agency could open its own decisions and judgment to greater public scrutiny and challenge to the extent that more data is disclosed and made available to the public.

https://doi.org/10.1017/9781316544587.003 Published online by Cambridge University Press

TABLE 2.3 *Comparison of state roles: genomics and clinical trials*

State Role	HGP	Post-HGP (NIH)	Private Sector	Clinical Trials
Convenor	Yes	Sometimes	Sometimes	No
Funder	Yes	Yes	No	Sometimes
Creator	Partially	Partially	No	No
Collaborator	Yes	Sometimes	Sometimes	Sometimes
Endorser	Yes	Yes	Sometimes	No
Curator	Yes	Yes, but declining	Sometimes	Limited
Regulator	Yes	Yes	Yes, if data is state-curated	Yes (FDA)
Enforcer	Yes (though never exercised)	Yes (though rarely exercised)	Yes, if data is state-curated	Yes, though not vigorously
Consumer	No	No	No	Yes (FDA)

has demonstrated a propensity for engaging in successful collaborative activity with private sector firms. This trend may be worth encouraging further in the area of clinical trials data.

The same may not be true with respect to the curatorial function. NIH through NCBI currently acts as the curator of summary clinical trials data submitted to ClinicalTrials.gov. Unlike NCBI's value-adding role with respect to genomic data, the agency adds little to clinical trials data, for the reasons discussed earlier. Thus, it is not clear that NCBI's substantial expertise is necessary to host and manage a clinical trials data commons. Particularly in view of federal budgetary constraints, it may be worth considering whether other options, such as outsourcing the curatorial function to a private sector contractor selected through competitive bidding, may result in greater efficiencies and cost savings.

Finally, as the previous examples illustrate, NIH's enforcement of rules relating to both genomics data and clinical trials data has been lackluster, at best. This failure of enforcement may arise from the close collaborative relationships between NIH and its funded researchers. To improve the effectiveness of rules enforcement, planners may wish to consider moving enforcement responsibilities away from NIH and to a different agency. With respect to clinical trials data, FDA may be a more logical choice, as it already exists in an arm's length, if not adversarial, relationship to the pharmaceutical and medical device firms that it regulates. It may also be possible to designate a different governmental agency to fill the enforcement role, either an existing agency more accustomed to policing and enforcement activity (such as the Federal Trade Commission or Department of Justice) or a new agency or subagency within NIH or FDA. Any of these approaches would sever the potential ties of loyalty and familiarity between the research-focused arm of NIH and the researchers whom it seeks to police.

CONCLUSIONS

This chapter has shown that the state plays a multiplicity of roles in the formation and management of large repositories of biomedical research data, extending well beyond the traditional big science model of the state as a creator/provisioner of data commons. The nine discrete state roles and the analytical framework described in this chapter offer a means by which state engagement with data-intensive research projects may be compared across agencies, fields, and national borders. This framework may be used to assess the effectiveness of state engagement in such research programs.

In particular, a number of lessons may be learned from NIH's evolving role in the genome commons, from funder and primary overseer of the HGP to convenor, collaborator, and curator. One such lesson suggests that the state may be a good curator of research data, whether governmental laboratories, grant-funded academic institutions, or the private sector generated that data. So long as data is intended to be disseminated to the public in a uniform manner, a state actor with requisite expertise may be the most logical candidate for that curation role. Nevertheless, extensive and expert data curation comes at a significant cost, and as the body of available scientific data continues to grow, the state's ability to offer curation services at no cost to the public may become strained.

Additional inefficiencies may arise from comingling the state's collaboration and enforcement roles. NIH's lackluster enforcement record as to both the genome commons and ClinicalTrials.gov suggests that alternative enforcement mechanisms should be considered for future data commons.

The state's engagement with the genome commons offers insights to planners of future research data commons, including the proposed clinical trials data commons. But while NIH's achievements in the area of genomics data should be applauded, they may not always translate directly to other research domains. For example, there may be more cost-effective or streamlined mechanisms for sharing research data that mandate less active curation and updating than the substantial NCBI resources devoted to genomic data.

In general, it is hoped that the analytical framework developed in this chapter will help researchers and policymakers configure state engagement in new data commons in a manner that will enhance the efficiency and effectiveness of data-sharing arrangements and improve overall social welfare.

REFERENCES

Ankeny, Rachel A. and Sabina Leonelli, Valuing Data in Postgenomic Biology: How Data Donation and Curation Practices Challenge the Scientific Publication System, in *Postgenomics: Perspectives on Biology after the Genome* 126 (Sarah S. Richardson and Hallam Stevens eds., Duke University Press 2015).

Benkler, Yochai, Review: Commons and Growth: The Essential Role of Open Commons in Market Economies, 80 *U. Chicago L. Rev.* 1499 (2013).

https://doi.org/10.1017/9781316544587.003 Published online by Cambridge University Press

Benson, Dennis, David J. Lipman, and Jams Ostell, GenBank, 21 *Nucleic Acids Research* 2963 (1993).
Bermuda Principles, Summary of Principles Agreed at the First International Strategy Meeting on Human Genome Sequencing (1996), www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml
Boyle, James, The Second Enclosure Movement and the Construction of the Public Domain, 66 *Law & Contemp. Probs.* (2003).
Bozeman, Barry, Technology Transfer and Public Policy: A Review of Research and Theory, 29 *Research Policy* 627 (2000).
ClinicalTrials.gov, Overview of FDAAA and Other Trial Registration (2012), https://prsinfo.clinicaltrials.gov/trainTrainer/Overview-FDAAA-Other-Regist-Policies.pdf
ClinicalTrials.gov, Trends, Charts, and Maps (2015), https://clinicaltrials.gov/ct2/resources/trends
Contreras, Jorge L., Constructing the Genome Commons in *Governing Knowledge Commons* 99 (Brett M. Frischmann, Michael J. Madison, and Katherine J. Strandburg eds., Oxford University Press 2014).
Contreras, Jorge L., Bermuda's Legacy: Policy, Patents, and the Design of the Genome Commons, 12 *Minn. J.L. Sci. & Tech.* 61 (2011).
Contreras, Jorge L. and Jerome H. Reichman, Sharing by Design: Data and Decentralized Commons, 350 *Science* 1312 (2015).
Cook-Deegan, Robert, *The Gene Wars – Science, Politics, and the Human Genome* (Norton 1994).
Couzin-Frankel, Jennifer, Trust Me, I'm a Medical Researcher, 347 *Science* 501 (2015).
Demaine, Linda J. and Aaron Xavier Fellmeth, Reinventing the Double Helix. A Novel and Nonobvious Reconceptualization of the Biotechnology Patent, 55 *Stanford Law Review* 303 (2002).
Demsetz, Harold, Toward a Theory of Property Rights, 57 *Am. Econ. Rev.* 347 (1967).
Ehrenfeld, David W., *Conserving Life on Earth* (Oxford University Press 1972).
Exec. Off. President, Big Data: Seizing Opportunities, Preserving Values (2014).
Food and Drug Administration Amendments Act of 2007 (FDAAA).
Fndn. for the Natl. Inst. Health (FNIH), How We Work (2015), http://fnih.org/about/how-we-work.
Frischmann, Brett M., *Infrastructure: The Social Value of Shared Resources* (Oxford University Press 2012).
Frischmann, Brett M., Two Enduring Lessons from Elinor Ostrom, 9 *J. Institutional Econ.* 387 (2013).
Galison, Peter, The Many Faces of Big Science in *Big Science: The Growth of Large-Scale Research* 1 (Peter Galison and Bruce Hevly eds., Stanford University Press 1992).
The GAIN Collaborative Research Group (GAIN), New Models of Collaboration in Genome-Wide Association Studies: The Genetic Association Information Network, 39 *Nature Genetics* 1045 (2007).
Hardin, Garrett, The Tragedy of the Commons, 162 *Science* 1243 (1968).
Heller, Michael A. and Rebecca S. Eisenberg, Can Patents Deter Innovation? The Anticommons in Biomedical Research, 280 *Science* 698 (1998).
Hess, Charlotte and Elinor Ostrom, Introduction: An Overview of the Knowledge Commons in *Understanding Knowledge as a Commons: From Theory to Practice* 4 (Charlotte Hess and Elinor Ostrom eds., MIT Press 2007).
Hilgartner, Stephen, Biomolecular Databases – New Communication Regimes for Biology? 17 *Science Communication* 240 (1995).

Holden, Arthur, The SNP Consortium: Summary of a Private Consortium Effort to Develop an Applied Map of the Human Genome, 32 *BioTechniques* 22–26 (2002).
Holden, Arthur L. et al., The International Serious Adverse Events Consortium, 13 *Nat. Rev. Drug Discovery* 795 (2014).
Holdren, John P., Director, Off. Sci. Tech. Policy, Memorandum for the Heads of Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Scientific Research (2013).
Holdren, John P., The 2017 Budget: Investing in American Innovation (2016), www.whitehouse.gov/sites/default/files/microsites/ostp/fy_17_ostp_slide_deck.pdf
Hudson, Kathy L. and Francis S. Collins, Sharing and Reporting the Results of Clinical Trials, 313 *J. Am. Med. Assn.* 355 (2014).
Inst. of Med. (IOM) and National Research Council, Large-Scale Biomedical Science (National Research Council 2003).
Inst. of Med. (IOM), Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk (National Research Council 2015).
Int'l HapMap Consortium, The International HapMap Project, 426 *Nature* 789 (2003).
Kaiser, Jocelyn, Celera to End Subscriptions and Give Data to Public GenBank, 308 *Science* 775 (2005).
Kaiser, Jocelyn, Funding for Key Data Resources in Jeopardy, 351 *Science* 14 (2016).
Loiter, Jeffrey M. and Vicky Norberg-Bohm, Technology Policy and Renewable Energy: Public Roles in the Development of New Energy Technologies, 37 *Energy Policy* 85 (1999).
Madison, Michael J., Knowledge Curation, 86 *Notre Dame L. Rev.* 1957 (2011).
Madison, Michael J., Brett M. Frischmann, and Katherine J. Strandburg, Constructing Commons in the Cultural Environment, 95 *Cornell L. Rev.* 657 (2010).
Mazzucato, Mariana, *The Entrepreneurial State: Debunking Public vs. Private Sector Myths* (Public Affairs 2013).
Mcelheny, Victor K., *Drawing the Map of Life – Inside the Human Genome Project* (Basic Books 2010).
Mishra, A., P. N. Schofield, and T. M. Bubela, Sustaining Large-Scale Infrastructure to Promote Pre-Competitive Biomedical Research: Lessons from Mouse Genomics, 33 *New Biotechnology* 280 (2016).
Nat'l. Acad. Sci., Eng. and Med. (NAS), Assessment of the Usefulness and Availability of NASA's Earth and Space Science Mission Data (2002), http://sites.nationalacademies.org/cs/groups/ssbsite/documents/webpage/ssb_051716.pdf.
Nat'l. Center for Advancing Translational Sciences (NCATS), Alliances at NCATS (2015), https://ncats.nih.gov/alliances/about.
Nat'l Center for Biotechnology Info. (NCBI), *The NCBI Handbook* (2nd ed.) (2013).
Nat'l Human Genome Research Inst. (NHGRI), NHGRI Policy Regarding Intellectual Property of Human Genomic Sequence (1996), www.genome.gov/10000926.
Nat'l. Inst. Health (NIH), Genomic Data Sharing (GDS) – Categories, Statistics, and Summary Information on Policy Violations (2015), https://gds.nih.gov/20ComplianceStatistics_dbGap.html.
Nat'l. Inst. Health (NIH), NIH Public-Private Partnership Program – Partnership Examples (2010), http://ppp.od.nih.gov/pppinfo/examples.asp.
Nat'l Inst. Health (NIH), Budget (2017), https://www.nih.gov/about-nih/what-we-do/budget
Nat'l Library Medicine (NLM), Congressional Justification FY 2016, www.nlm.nih.gov/about/2016CJ.html#Justification.
Nat'l. Research Council (NRC), Rising to the Challenge: US Innovation Policy for the Global Economy (National Research Council 2014).

Nat'l. Research Council (NRC), Reaping the Benefits of Genomic and Proteomic Research (National Research Council: 2006), www.nap.edu/catalog/798/infrastructure-for-the-21st-century-framework-for-a-research-agenda

OECD, Data-Driven Innovation: Big Data for Growth and Well-Being (OECD Publishing 2015).

Ostrom, Elinor, *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge University Press 1990).

Ostrom, Elinor, *Understanding Institutional Diversity* (Princeton University Press 2005).

Paltoo, Dina N. et al., Data Use under the NIH GWAS Data Sharing Policy and Future Directions, 46 *Nature Genetics* 934 (2014).

Rai, Arti Kaur, Regulating Scientific Research: Intellectual Property Rights and the Norms of Science, 94 *Northwestern U. L. Rev.* 77 (1999).

Rai, Arti K. and Rebecca S. Eisenberg, Bayh-Dole Reform and the Progress of Biomedicine, 66 *Law & Contemp. Probs.* 289 (2003).

Reichman, Jerome H., Paul F. Uhlir, and Tom Dedeurwaerdere, *Governing Digitally Integrated Genetic Resources, Data, and Literature: Globally Intellectual Property Strategies for a Redesigned Microbial Research Commons* (Cambridge University Press 2016).

Rhodes, Catherine, *Governance of Genetic Resources: A Guide to Navigating the Complex Global Landscape* (Edward Elgar 2013).

Rose, Carol M., The Comedy of the Commons: Commerce, Custom, and Inherently Public Property, 53 *U. Chicago L. Rev.* 711 (1986).

Saito, Hiroki and Christopher J. Gill, How Frequently Do the Results from Completed US Clinical Trials Enter the Public Domain? A Statistical Analysis of the Clinicaltrials.Gov Database, 9 *PLoS One* (2014).

Sarnoff, Joshua D., Government Choices in Innovation Funding (With Reference to Climate Change), 62 *Emory L. Rev.* 1087 (2013).

Scotchmer, Suzanne, *Innovation and Incentives* (MIT Press 2004).

Shreeve, James, *The Genome War* (Knopf 2004).

Smith, Robert J., Resolving the Tragedy of the Commons by Creating Private Property Rights in Wildlife, 1 *CATO J.* 439 (1981).

Strandburg, Katherine J., Brett M. Frischmann, and Can Cui, The Rare Diseases Clinical Research Network and the Urea Cycle Disorders Consortium as Nested Knowledge Commons, in *Governing Knowledge Commons* 155 (Brett M. Frischmann, Michael J. Madison, and Katherine J. Strandburg eds., Oxford University Press 2014).

Strasser, Bruno J., GenBank – Natural History in the 21st Century? 322 *Science* 537 (2008).

US Food & Drug Admin., *Critical Path Initiative* (2017), http://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/.

US Food and Drug Admin., FDA News Release: FDA and International Serious Adverse Events Consortium Complete Third Data Release (2010), www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm201286.htm.

US Patent and Trademark Office, Revised Utility Examination Guidelines; Request for Comments, 64 *Fed. Reg.* 71440 (1999).

US Securities and Exchange Comm., Report to the Congress on Section 107 of the National Securities Markets Improvement Act of 1996 – Privatization of EDGAR (1997), https://www.sec.gov/news/studies/edgpriv.htm.

Vertinsky, Liza S., Patents, Partnerships, and the Pre-Competitive Collaboration Myth in Pharmaceutical Innovation, 48 *U.C. Davis L. Rev.* 101 (2015).

White House, Exec. Order 13642, Making Open and Machine Readable the New Default for Government Information (2013).