Promoting Public Good Uses of Scientific Data: A Contractually Reconstructed Commons for Science and Innovation

by

J. H. Reichman and Paul F. Uhlir*

Abstract

[to be completed]

Table of Contents

* J.H. Reichman is the Bunyan S. Womble Professor of Law at Duke Law School; Paul F. Uhlir is Director of International Scientific and Technical Information Programs at the National Research Council, Washington, D.C. The opinions expressed in this paper are the authors' and not necessarily those of the National Research Council

A. An E-Commons for Science

1. The basic concept

2. Differentiating centralized from decentralized suppliers

B. Implementing the E-Commons Approach

1. Instituting an unconditional public domain

2. Conditional public domain mechanisms

a. Characteristics of an impure domain

b. Sectoral evaluations

(i) The public sector

(ii) The academic sector

Linking the communities

Compensatory liability

Administrative considerations

(iii) The private sector

Introduction

[Introductory section to be written later]

I. DECONSTRUCTING THE PUBLIC DOMAIN IN SCIENTIFIC DATA

A. Defining the Public Domain

The idea of a public domain in information arises from an imperfect analogy to the concept of a commons, or publicly owned or managed land, in real property law.[1] Yet, until recently, there have been few scholarly writings that have explored the public domain in the information context[2] and none, to our knowledge, that has defined or examined the public domain in scientific data[3] in any comprehensive manner.

For the purposes of this paper, we define the public domain in terms of sources and types of information whose uses are not impeded by legal monopolies grounded in statutory intellectual property regimes, and which is accordingly available to some or all members of the public without authorization. For analytical purposes, the public domain in information, including especially scientific and technical (S&T) data, may be subdivided into three major categories:
(1) information that is not subject to protection under exclusive intellectual property rights;

[1] See, e.g., Elizabeth Longworth (1999). However, unlike the commons in rural England that had to be carefully managed to prevent their destruction fromoveruse, an intangible information commons has public good characteristics that make it non-depletable (see the last section in Part I).
[2] *See, e.g.*, Lange (1981), Litman (1990), Samuels (1993), and Benkler (1999).
[3] [cites]

(2) information that qualifies as protectable subject matter under some intellectual property regime, but that is contractually designated as unprotected;[4] and (3) information that becomes available under statutorily created immunities and exceptions from proprietary rights in otherwise protected material, such as the "fair use" exception in copyright law[5], which promote certain public-interest goals at the expense of proprietors' exclusive rights.

1. Information Not Subject to Legal Monopolies

Three subsets of information fall within this category. The first consists of information that intellectual property rights cannot protect because of the nature of the source that produced it. The second comprises otherwise protectable information that has lapsed into the public domain because its statutory term of protection has expired. The third includes ineligible or unprotectable components of otherwise protectable subject matter.

---

[4]*See infra* text accompanying notes _____. 17 USC §107. Another example would be the research exemption in patent law, 35 USC § ____, which is narrowly construed in the U.S. and more broadly construed in the E.U. *See, e.g.*, Rai & Eisenberg (this conference). Eisenberg (Norms of Science). For the sake of economy, this paper will focus on legal regimes that directly confer exclusive property rights on collections of information as such, and it will not examine patent law except at the margins.

[5]For present purposes, moreover, we tend to ignore liability rules, especially trade secret law, which confers no exclusive property rights in confidential information and permits reverse-engineering by honest means, and as well as unfair competition law, which interdicts market-destructive conduct. *See* RESTATEMENT THIRD OF TORTS. These regimes limit access to the public domain without impoverishing it (except, of course, where, say, reverse-engineering is infeasible). *See e.g.*, Pamela Samuelson, Reverse Engineering (2001). On the whole, however, the critical problem today is the inability of traditional liability rules to provide costly information products with natural lead time, which results in exaggerated claims of market failure and in a proliferation of *sui generis* exclusive property rights in small-scale applications of information (*qua* know-how) to industry. *See generally* J. H. Reichman , Legal Hybrids Between the Patent and Copyright Paradigms , __ COLUMB. L. REV. ___ (1994); J. H. Reichman, Collapse of the Patent Copyright Dichotomy: Premises for a Restructured International Intellectual Property System, ___ CARDOZO J. LAW & ARTS ____, (1995).

a) Information that cannot be protected because of its source

The U.S. government is by far the largest creator, user, and disseminator of data and information in the world.[6] Most of the material produced by both federal and state governments cannot be legally protected. To this end, the 1976 Copyright Act prohibits the federal government from claiming copyright protection of the information it produces.[7]

There are a number of well-established reasons for this policy. The government needs no legal incentives to create the information; the taxpayer has already paid once for the production of a database or report and should not pay twice; transparency of governance and democratic values would be undermined by limiting broad dissemination and use of public data and information; citizens' First Amendment rights might be compromised; and the nation generally benefits in myriad ways from broad, unfettered access to and use of government databases and other public information by all citizens to promote economic, educational, and cultural values.[8] It is primarily the latter justification, which encompasses the value of public-domain scientific and technical data for the conduct of research and our national system of innovation in particular,[9] that is the focus of discussion in this paper.

The existing situation with regard to legal protection of databases and other productions by state and local governments is not as straightforward as in the federal context. Section 105 of the 1976 Copyright Act does not expressly ban copyright claims in the works of non-federal government entities. Many states have nonetheless enacted open records laws that prohibit protection of government information, encourage open dissemination to the public, and contain provisions analogous to the Freedom of Information Act (FOIA).[10] There is no uniformity among the states in these areas, however, and there are many exceptions that allow state and local jurisdictions to protect some types

[6] Weiss and Backlund (1997), NRC (1999)

[7] 17 U.S.C., section 105, which states: "Copyright protection under this title is not available for any work of the United States Government."

[8] Weiss and Backlund (1997), NRC (1999), OTA (1986).

[9] *See, e.g.*, Nelson, ed., NATIONAL SYSTEMS OF INNOVATION.

[10]

of information generated by selected agencies, even in those states that have enacted open records laws. Consequently, many state and local agencies do currently protect their databases and other productions under copyright and contract laws, and these agencies would likely make use of any additional protection that new federal or state laws might provide. Nevertheless, we believe that the public-policy arguments that justify non-protection in the federal context should in principle apply to information produced by state and local governments.

The federal government also produces the largest body of public-domain data and information used in scientific research and education, both in terms of the volume as well as in terms of the percentage of material produced. For example, the U.S. federal government alone spends more than $80B[11] on its research programs, with a significant percentage of that invested in the production of primary data sources; in higher-level processed data products, statistics, and models; and in S&T information, such as government reports, technical papers, research articles, memoranda, and other such analytical material.[12] The bulk of the data and information thus produced in government programs automatically enters the public domain, year after year, with no proprietary restrictions (although the sources are not always easy to find!), with the important exception of certain limitations on access for reasons of national security,[13] or protection

[11] DOE (2000). See also NRC (1999).

[12] A very preliminary OECD estimate of the percentage of public research investment that supports the creation of scientific and technical data and information places it in the range of 50 to 80 percent (personal communication, OECD, 2001).

[13] Cite legis. Several U.S. laws and policies based on national security concerns recently have been adopted or proposed to limit the scope of data and information that can be published, discussed openly at venues that include foreign nationals, or transferred internationally. The tightening of export control restrictions under the International Trade in Armaments Regulations (ITAR) has placed limits on scientists and engineers in academic discourse, publishing, and data sharing, particularly in the area of civil space systems, such as space science and environmental remote sensing, GPS, and communications satellites. This has been accompanied by the introduction of a new category of quasi-classified information known as "Sensitive Unclassified Technical Information" (SUTI), which has been used to exercise prior restraints of dubious constitutional validity on the disclosure of such information or on the free association of U.S. scientists with foreign colleagues.

of personal privacy or confidentiality.[14] These various limitations on the public domain accessibility of federal government information, while often justified, must nonetheless be balanced against the rights and needs of citizens to access and use it.

The advent of the era of "big science" following World War II established a framework for the planning and management of large-scale basic and applied research programs.[15] Most such research was conducted in the physical sciences and engineering, fueled largely by the Cold War and related defense requirements. Although a substantial portion of this research was classified, at least initially, most of the government and government-funded research results that these programs generated entered the public

---

In another related development, legislation similar to the British "official secrets act" was passed by Congress in 2000 that would have allowed administrative censorship and a broader use of "national security" reasons to withhold public information. Its enactment was narrowly averted by a veto by President Clinton last fall only after civil libertarians expressed substantial concerns. The same legislation has been introduced this year.

Finally, recent incidents of espionage have led to greatly increased restrictions on scientists working in government laboratories that conduct classified research, further diminishing the potential scope for publication and communication of public-domain data and information. The tightening of disclosure of classified and "sensitive" government data and information based on national security concerns is likely to intensify, at least in the near term, in reaction to the recent terrorist attacks. (add cites)

[14] Cite legis. The protection of individual privacy and the related confidentiality of private information is another countervailing legal and ethical value that is used to limit the availability and dissemination of information otherwise in the public domain. Examples of well-established applications of this exception include primary census data and the data on individual subjects in biomedical research. As various forms of research become ever-more intrusive and revealing, however, the privacy/confidentiality exception is taking on added importance and scope. For instance, research involving individuals' genetic information or genetic testing of patients in government service, has led to the adoption of stricter laws on limiting the disclosure of such data or results to unauthorized third parties. In the area of environmental remote sensing, certain satellite or ground-based observations recorded by the government are kept confidential in order to protect the privacy of individuals, or to withhold the precise location of endangered species from prospective poachers. (add cites)

[15] See *Big Science*

domain. This research model yielded a succession of spectacular scientific and technological breakthroughs and well-documented socioeconomic benefits.[16]

The hallmark of big science, more recently referred to as "megascience,"[17] has been the use of large research facilities or research centers and of "facility-class" instruments, which are most usefully characterized as observational and experimental.[18] In the observational sciences, some of the most significant advances initially occurred in the space and earth sciences as offshoots of classified military and intelligence space technologies and NASA's Apollo program. Notable examples of large observational facilities have included space science satellites for robotic solar system exploration, ground-based astronomical telescopes, earth observation satellites, networks of terrestrial sensors for continuous global environmental observations and global change studies, and, more recently, automated genome decoding machines.[19] Major examples in the experimental sciences have included facilities for neutron beam and synchotron radiation sources, large lasers, supercolliders for high-energy particle physics, high-field magnet laboratories, and nuclear fusion experiments.[20]

The data from many of these government and government-funded research projects have been openly shared and archived in public repositories. Hundreds of specialized data centers have been established by the federal science agencies or at universities under government contract. A few well-known examples of the government's public-domain data archiving and dissemination activities include the NASA Space Science Data Center,[21] the National Oceanic and Atmospheric Administration's (NOAA) National

---

[16] John A. Armstrong (October 13, 1993), "Is Basic Research a Luxury that Our Society Can No Longer Afford?," Karl Taylor Compton Lecture, Massachusetts Institute of Technology.; and other cites.

[17] See, e.g., the OECD Megascience Forum Web site at http://www.oecd.org/dsti/mega/.

[18] See Bits of Power (1997), at 58-61.

[19] Id.

[20] Id.

[21] [provide URL]

Data Centers,[22] the U.S. Geological Survey's Earth Resources Observation Systems (EROS) Data Center, and the National Center for Biotechnology Information at the National Institutes of Health.[23]

The situation with regard to basic, primary data has been different in "small science," that is to say, individual-investigator driven research, which remains the dominant form of practice in many scientific areas, both experimental and observational. In the experimental or laboratory sciences such as chemistry, or behavioral or biomedical research, researchers use large databases to a much lesser extent for advancement, depending instead on the use of individual, repeatable experiments or observations.[24] Instead of raw observational data, the laboratory sciences rely on the use of highly evaluated data sets and on the published scientific literature. Because of the extremely specialized, labor-intensive nature of evaluated data sets, many are produced outside government and made available in proprietary publications or databases. Nevertheless, some public-domain government sources exist for these types of data, even though they are smaller in number and volume than the sources of observational data.[25]

The "small science," independent-investigator approach also characterizes a large area of field work and studies, such as biodiversity, ecology, microbiology, soil science, and anthropology. Here, too, many individual or small-team data sets or samples are collected and analyzed independently.[26] Traditionally, the data from such studies have been extremely heterogeneous and unstandardized, with few of the individual data holdings deposited in public data repositories or even openly shared.

The widespread use of digital computing that began in the 1980s, and especially the establishment of the World Wide Web in the early 1990s, has led to exponential

---

[22] [provide URLs]
[23] [provide URL]
[24] NRC (1995a).
[25] See Bits of Power, supra note __, at __.
[26] See, NRC (1995b).

increases in the amount of digital data and information created in all sectors, not least in government research. These advances have given rise to both quantitative and qualitative changes in the production, dissemination, and use of scientific data, and they have changed the way that science itself is conducted.[27] Entirely new tools and techniques have been developed, such as the use of massively parallel supercomputing (for large database modeling and analysis), data mining, data animation and visualization, geographic information systems (GIS) for the integration of spatial data, collaboratories for the conduct of virtual experimentation, and computer-aided design and manufacturing, among many others. One of the most significant changes affecting the availability and dissemination of federal materials is that science agency Web sites now typically permit both direct and indirect access to their own and other related public-domain data and information resources. Almost all such data are free and available to anyone with access to an Internet connection anywhere in the world.

Moreover, the rise of digitally networked information, coupled with the development of sophisticated data management tools and techniques, has made it possible for the previously unconnected, individual-investigator driven, "small science" fields of research to become fully interconnected, collaborative, and more open with their specialized data sources. These prospects have prompted a reorganization of some previously small science fields, such as genomic studies in molecular biology, into big science programs, such as the Human Genome Project, and have led to the establishment of "bioinformatics" as a new organizing principle in biology.

Scientific and other kinds of data and information generated by the governments of other nations may also end up in the public domain and become available internationally,[28] but generally the quantities are much smaller than the information resources generated by the U.S. government, both in terms of the total amount and as a percentage of the total, and the public-access policies are much less certain than those applicable in the U.S. Notable examples of foreign sources of public-domain data are the World Data

[27] Id., at __. See also, A QUESTION OF BALANCE, and other cites.
[28] See EC Green Paper, 1999

Centers for geophysical, environmental, and space data[29] and the human genome databases in Europe and Japan.[30] However, a key issue for both the exploitation of public data resources and for cooperative research generally is the asymmetry between the United States and foreign government approaches to the public-domain availability of scientific data.[31]

b) Information whose term of statutory protection has expired

Under United States copyright law, the term of protection is long—the life of the author plus 70 years, or for works made for hire, the shorter of 95 years from first publication or 120 years from the date of creation.[32] Other nations grant similar terms of protection,[33] in accord with the international minimum standard under the Agreement on Trade Related Aspects of Intellectual Property Law (TRIPS Agreement).[34] This, too, constitutes an enormous body of freely available literature and information with great cultural and historical significance.

Some materials in this category have obvious relevance to certain types of research, especially in the social sciences and the humanities. Even some of the "hard" sciences can derive substantial value from public-domain data and information that are decades or even many centuries old. For example, the extraction of environmental information from a broad range of historical sources can help establish climatological trends, or assist in identifying or better understanding a broad range of natural phenomena.[35] Ancient Chinese writings are proving useful in identifying herbal medicines

---

[29] Cite NRC reports on WDC system and provide URLs

[30] See the European Molecular Biology Laboratory Web site at http://www.ncbi.edu, and the DNA Database of Japan Web site at http://www.ddjb.nig.ad.jp (check URLs)

[31] *See infra* text accompanying notes _____.

[32] 17 U.S.C., section 302

[33] (add cites), E.U. Directive.

[34] *See* TRIPS Agreement, *supra* note ___, arts. _____.

[35]

for modern pharmaceutical development, and proposals for more systematic and ambitious databases concerning traditional know-how and medicines are on the table.[36] Nevertheless, because of the long lag time before entering the public domain, most of the information in this category lacks relevance to most types of state-of-the-art research.

c) Ineligible subject matter or unprotectable components of eligible subject matter

Copyright law protects only original and creative works of authorship,[37] and its scope of protection extends only to the expressive content embodied in original works that fall within the codified list of eligible subject-matter categories.[38] Facts as such are excluded, although compilations of facts that evince a modicum of creative selection and arrangement are copyrightable.[39] Also ineligible in any "idea, procedure, process, system, method of operation, concept, principle or discovery" incorporated into an otherwise copyrightable work.[40]

In principle, this combination of rules excludes protection for random and complete assortments of data that lack sufficiently original and creative criteria of selection and arrangement.[41] Even when such criteria exist, and the relevant compilation qualifies as an eligible work of authorship, the Supreme Count has ruled that copyright protection does not extend to either the ideas or disparate facts set out in the work and which may be used freely.[42]

---

[36] Cite presentation given at bilateral U.S.-China CODATA data symposium in Beijing (2000); Discussion of Correa paper at WTO; WIPO projects.

[37] 17 U.S.C. §102(a); Feist Publications, Inc. v. Rural Telephone Service Co.,499 U.S.340 (1991).

[38] 17 U.S.C. §§102(a), 103; Feist v. Rural.

[39] *See supra* note __; *see also* _____.

[40] 17 U.S.C. section 102(b).

[41] *See, e.g.,* Reichman & Uhlir (1999) (citing authorities).

[42] Feist v. Rural (1991). Note, however, that some courts have stretched copyright law to protect some "methods" of compilation, *see infra* notes _____ and accompanying text, and some methods ineligible for copyright protection, including business methods. *See, e.g.,* State Street Bank; Rochelle Cooper Dreyfus, Unfair Competition rules may also sometimes be invoked against the wholesale

One of the largest categories of scientific information in the United States consists of ineligible collections of data or the noncopyrightable contents of otherwise copyrightable works, including databases, articles, or reference books. This category of public-domain information, while highly distributed among all types of proprietary works, plays a fundamental role in supporting research and education, especially in the data-intensive sciences.[43] However, strenuous efforts are being made to devise new forms of protection for all of this previously unprotectable subject matter, as explained later in this paper.[44]

2. Information Expressly Designated as Unprotected

A second major source of public-domain information is that which is created in the academic and private sectors, typically with government funding, and that has been contractually designated as unprotected. Such information, especially in the form of scientific data sets or more elaborately prepared databases, is made freely available for others to use, frequently through deposit in government or university data centers or archives.[45] Less frequent, but nonetheless important, examples are found in proprietary information created by industry, such as old oil exploration data sets with potentially significant geophysical research applications, which are subsequently donated to data centers or archives for open and unrestricted dissemination.[46] Databases and other information produced in academic settings, in not-for-profit institutions, or in industry, will become presumptively protectable under any available legal regime, however, unless such material is expressly placed in the public domain. The public domain in this case must be actively created, rather than passively conferred.

---

duplication of uncopyrightable compilations. *See, e.g.,*INS v. AP; NBA v. Motorola; Reichman & Samuelson (1997).

[43]*See, e.g.,* Reichman & Uhlir (citing authorities).

[44]*See infra* text accompanying notes _____.

[45][cites]

[46] R. Stephen Berry (2001), "Is electronic publishing being used in the best interests of science? The scientist's view," presentation given at the Second ICSU-UNESCO International Conference on Electronic Publishing, Paris, France, pp. 1-2.

Much like government scientists, academic researchers typically are not driven by the same motivations as their counterparts in industry and publishing. Public-interest research is not dependent on the maximization of profits and value to shareholders through the protection of proprietary rights in information; rather, the motivations of government and not-for-profit scientists are predominantly rooted in intellectual curiosity, the desire to create new knowledge, peer recognition and career advancement, and the promotion of the public interest. As the Home Secretary for the National Academy of Sciences, Stephen Berry, recently noted:

> Scientists are not, for the most part, motivated to do research in order to make money. If they were, they would be in different fields. The primary motivation for most research scientists is the desire for influence and impact on the thinking of others about the natural world—unless the desire for their own personal understanding is even stronger…. The currency of the researcher is the extent to which her or his ideas influence the thinking of others…. What this implies is that the distribution of the results of research has an extremely high priority for any working scientists, apart from those whose work is behind proprietary walls.

These values and goals are best served by the maximum availability and distribution of the research results, at the lowest possible cost, with the fewest restrictions on use, and the promotion of the reuse and integration of the fruits of existing results in new research. The public domain in S&T databases, and the long-established policy of full and open access to such resources in the government and academic sectors,[47] reflects these values and serves these goals.

The policy of "full and open" access or exchange has been defined in various U.S. government policy documents and in NRC reports as "data and information derived from publicly funded research are [to be] made available with as few restrictions as possible, on a nondiscriminatory basis, for no more than the cost of reproduction and distribution" (that

[47] *See infra* note ____.

is, the marginal cost of delivery).[48] This policy is promoted by the U.S. government in academia, with varying degrees of success, and in most cooperative research, whether in large, institutionalized research programs, such as global change studies or the human genome project, or in smaller-scale collaborations involving individual investigators.

The norms and practices governing the placement of such information in the public domain nonetheless tend to be specific to a discipline, institution, or research program, and vary significantly even within the United States.[49] In general, the rationale for extending available forms to legal protection to, say, databases gathered, organized, or maintained by academics has seemed weak or *de minimis*, as distinct from commercial investors who may risk substantial resources of their own in such activities. At the same time, other competing public policy reasons may not support an automatic exclusion of some nonprofit entities from availing themselves of legal protection, despite the absence of risk aversion and related rationales.

On the one hand, there are strong arguments for denying grantees of government funds the right to privatize their research results. There are both written and unwritten rules in most areas of basic research in academia, and even in government-funded basic research within industry, that the data collected or generated by grantees will be openly shared with other researchers, at least following some specified period of exclusive use–typically limited to 6 or 12 months–or until the time of publication of the research results based on those data.[50] This relatively brief period is intended to give the grantee sufficient time to organize, document, verify, and analyze the data being used in preparation of a research article or report for scholarly publication. Upon publication, or at the expiry of the specified period of exclusive use, the data in many cases are placed in a public

---

[48] *Bits of Power*, supra note __, at 15-16. See also National Research Council, *On the Full and Open Exchange of Scientific Data* (1995), National Academy Press, Washington, D.C.

[49] For a partial inventory of U.S. science agency and other scientific institution policies with regard to open data availability requirements, see Paul Wouters, 5 October 2001, "A Web Scan of Existing Rules on Data-Sharing in US Research Funding Agencies," NIWI Research, prepared for the OECD.

[50] See *Bits of Power* (1997), at 79.

archive or Web site and expressly designated as free from legal protection, or they are made available directly by the researcher to anyone who requests access. The motivations and values that drive this not-for-profit research and educational activity are outside the sphere of commerce, as discussed above.

On the other hand, federal research policy encourages the commercialization, economic exploitation, and intellectual property protection of some fruits of academic research in certain circumstances. For example, the exclusion of works by federal government employees, within the scope of their employment, from copyright protection does not extend to grantees or contractors, who are allowed to copyright their research results. For the past 20 years, moreover, the Bayh-Dole Act has encouraged researchers who receive federal grants, and the universities that employ them, to patent the inventions arising from their federally-funded research. To date, the rules favoring the open sharing of upstream, unpatentable and noncopyrightable data flows, including S&T data derived from such research, and their placement in the public domain have generally trumped these countervailing policies and interests.

Although the cooperative and sharing ethos of science and the policy of full and open access that implements it, like all ideals, have never been fully realized, at least not across all of science, most scientists engaged in public-interest research do take the availability of both data and ideas for granted, much as we take air and water for granted. U.S. government-supported large facility-based research, in particular, has operated in a world in which there have been no exclusive rights in the data collected and used. As we discuss in Part II, however, this situation appears to be rapidly shifting to one that is or may become much more dominated by privatized and commercialized data activities.

3. Codified immunities and exceptions from proprietary rights

A final category of what may be considered public-domain information consists of statutorily created exceptions from applicable intellectual regimes, such as the exceptions

from copyright protection that favor teaching, research, and other educational activities,[51] the private use exception in E.U. law,[52] and the "fair use" exception in U.S. copyright law.[53] In this category, certain unprotected uses may be made of otherwise protected content under limited circumstances to advance the public interest in certain privileged policy goals.[54] Compulsory licenses may sometimes be enacted to promote these same goals, in lieu of an outright exception.[55] In either case, the theory is that the state may extract certain public-good concessions from proprietors of intellectual property rights in exchange for its willingness to enforce portable legal fences around intangible creations that would otherwise remain freely available owing to their nonrivalrous, ubiquitous, and inexhaustible character.[56]

In United States copyright law, considerable emphasis has been placed on the fair use exception to copyright protection,[57] which on a case-by-case basis may sometimes permit so-called "transformative" uses of otherwise protective information,[58] especially for such purposes as illustration, teaching, verification, and news reporting.[59] The strength of this exception varies with judicial attitudes, from period to period, and its consistency with international intellectual property law has been called into question.[60]

Because many so-called fair uses are allowed only in the context of not-for-profit research or education, this category of "public-domain uses," though relatively small, is especially important in the research context. It also tends to be the most controversial area

---

[51]*See, e.g.*, 17 U.S.C. §110(1)-(4).
[52][cites]
[53]17 U.S.C. §107. *See also* 35 U.S.C. §_____ (research exception in patent law); *supra* note ___.
[54]*See, e.g.*, [Supreme Court's parody decision]; *but see* [Supreme Court's decision in the Nation case].
[55]*See, e.g.*, Berne Convention, arts. _____ and Appendix (compulsory licenses for translations and other uses of scientific works in developing countries).
[56]*See, e.g.*, Kreiss; other cites. However, the affected class of proprietors will typically respond that it should not be made to subsidize the privileged activities in question.
[57]*See supra* note _____; DIGITAL DILEMMA.
[58]*See e.g.*, [Gone with the Wind case].
[59]*See* 17 U.S.C. §107 (preambular uses).
[60]*See, e.g.*, Ruth Gana Okediji, _____.

and is frequently in dispute by rights holders. Recently, however, courts have tended not to allow the fair use exception when technical means to avoid market failure are shown to exist.[61] Moreover, although copyright law has not typically associated fair uses and other exceptions with the "public domain" per se, a number of traditionally practiced immunities and exceptions, including fair use, may be construed as functional equivalents of public-domain uses, especially where science and education are concerned. The trend in changes to existing law, as well as in new *sui generis* "intellectual property" rights, is to severely curtail the scope of fair use and other exceptions for science and for other public-interest uses.[62]

C. The Economic Role and Value of the Public Domain in Scientific Data

In defining the nature of the public domain in scientific data, we have observed that the culture and process of science and innovation have become increasingly dependent upon the open availability and unrestricted use of data resources. Compelling economic principles support the continued existence of a vigorous public domain in scientific data, and there is a strong case for the exponential value that open and unrestricted data flows add to the economy and to society generally.

To better understand the value of the public domain in scientific data, one must distinguish between the respective roles of the public and private sectors in the development and dissemination of information products and services, generally, and of scientific data, specifically. As in all the mixed economies of the developed world, both the government and the private sector play a substantial role in the U.S. economy, although historically the government has performed a much lesser function in this country than in

---

[61] *See, e.g.*, American Geophysical (2nd Cir.); Wendy J. Gordon, Fair Use as Market Failure, _____ COLUM. L. REV. _____ ( ).

[62] *See infra* text accompanying notes _____.

other developed countries, largely limited to correcting the imperfections in private production.[63]

The broad limitations on the scope of U.S. government activity also prohibit the government from directly commercializing the information it produces and from competing with the private sector.[64] This constraint is also said to justify the prohibition on government from claiming intellectual property protection for the information it produces, and for requiring the placement of that information in the public domain, with a view to its broader exploitation by the private sector and all citizens. Other regulations prevent the federal government from pricing its information at a level greater than the incremental cost of dissemination, which excludes recouping the costs of producing that information, much less making a profit.[65] Indeed, the regulatory bias that favors charging no more than the marginal cost of disseminating the information,[66] means that, on the Internet, the price is zero. This policy differs from that of most other developed countries, where government or quasi-government agencies themselves may commercially exploit public information at commercial rates, and may also invoke the protection of that information under intellectual property law.[67]

The prohibition of the United States government's direct commercialization of its own information still begs the question of what types of information the government should

---

[63] Stiglitz, et al. (2000)

[64] OMB Circular A-76

[65] OMB Circular A-130 (1993), as codified in the Paperwork Reduction Act of 1995, 44 U.S.C. Part 35.

[66] Id., at __.

[67] For an overview of information policy in Europe and comparisons with U.S. information policy, see the Commission of the European Communities Green Paper (1999), *Public Sector Information: A Key Resource for Europe* and PIRA International (2000), *Commercial Exploitation of Europe's Public Sector Information*, Report for the European Commission, Directorate General for the Information Society. For a comparison of U.S. and E.U. policies with regard to public data production and dissemination in the area of meteorological data, see Pluijmers and Weiss, *Borders in Cyberspace: Conflicting Government Information Policies and Their Economic Impacts* (publication pending).

produce and make available via the public domain. Stiglitz, et al. have posited a number of rationales that conceivably justify a government in undertaking economic activity, of which two are particularly relevant to basic scientific data—the provision of public goods and the promotion of positive externalities.[68]

A public good has two essential characteristics that distinguish it from a private good: it must be *nonrivalrous* and *nonexcludable*.[69] Nonrivalrous means that there is no additional cost incurred in providing the good to an additional person (i.e., it has a zero marginal cost). Nonexcludable means that one cannot exclude others from deriving benefit from the good once it has been produced. There are, in fact, few public goods that fully satisfy both criteria, with national defense and the operation of lighthouses the frequently cited examples.

Information, particularly in its intangible form, also has public-good qualities.[70] An idea or a fact, once divulged, costs nothing to propagate and becomes impossible to keep from others. However, once information or a collection of facts housed in a database become fixed in a tangible medium, whether on paper or in digital form, they forfeit their exclusively public-good qualities. Embodied information can be treated as a private good, potentially excludable through intellectual property rights and physical forms of protection, and access to it can be traded for payment.[71] In this state, information becomes a "mixed good," or a quasi-public good, having only limited aspects of both public-good characteristics.[72] Despite the fact that information made available online still retains its nonrivalrous qualities, it can nonetheless be made excludable by using digital rights

[68] Joseph E. Stiglitz, Orszag, Peter R., and Orszag, Jonathan M. (2000), *The Role of Government in a Digital Age*, Computer and Communications Industry Association, Washington, D.C., pp. 31-35, citing Joseph E. Stiglitz (1988), *Economics of the Private Sector*, pp. 198-212.
[69] Id., at 32. See also ___.
[70] *See, e.g.,* Reichman & Franklin (1999) (dual functions of information); Eleonore Ostrum (this conference)(focusing attention on this phenomenon).
[71] Kreiss, Access.....; *see also*, McGowan.
[72] Pluijmers and Weiss (publication pending).

management technologies and contracts[73] (or through the tight control of access to the Internet in a totalitarian regime).

Basic, or fundamental, research is another activity that yields primarily a public good.[74] A new discovery in nature, the incremental advancement of an idea, or the observation of a natural phenomenon or event in the course of scientific research can be both nonrivalrous and nonexcludable. Even the collection of raw observations or facts in a database largely retains this public-good character. It is only when the fruits of such research are formed into economically valuable products and applications that they acquire mixed-good or private-good qualities.

Scientific data frequently partake of the public-good characteristics that reside in both information and basic scientific research. As the economy becomes increasingly information-based and science becomes much more data driven, there is an inherent implication that our traditional preference for all economic activity to be undertaken in the private sector may not be the optimal mode of organization in certain specific areas.[75] The creation and dissemination of scientific data become especially appropriate for consideration as governmental, or government-funded, activities.

Another potential justification for increasing government activity in the nation's economy that has particular relevance to the public domain in scientific data is the promotion of positive externalities. An externality may be defined as the action of one entity affecting the well-being of another, without appropriate compensation. A negative externality is the imposition of additional costs by entity A (for example, through the deleterious effects of pollution created by A) on entity B, without A's having to pay for those costs. Conversely, a positive externality confers benefits (e.g., technology) from A

---

[73] *See* Reichman & Franklin, *supra* note _____ , at _____ (discussing online restoration of the power of the two-party deal that was lost when the printing press was invented).

[74] Basic, or fundamental, research may be defined as research that leads to new understanding of how nature works and how its many facets are interconnected. *See* Armstrong (1993).

[75] Stiglitz, et al., at 40-41.

to B without full compensation to A.[76] Basic research, together with the creation and dissemination of scientific databases, especially in their raw form, may have no immediate economic applications or market, but they can lead subsequently to perhaps unanticipated or serendipitous advances and to whole new spheres of commerce. Such activities are prime examples of positive externalities that direct government support can greatly promote and that may not be undertaken at all without such support.

A related concept is a network externality, which arises when the value of using a particular type of product depends on the number of users.[77] Examples of products with high positive feedback from such network externalities include telephones and fax machines, if there are many users rather than only a few. Perhaps the quintessential product with positive network externalities is the Internet. A scientific database or other collections of information can potentially add a lot more value to society at large and to the economy if they are openly available on the Internet (assuming that production remains feasible in the absence of appropriability as would occur with government funding).

Indeed, the value of scientific data lies in their use.[78] Scientists were the pioneers of the Internet revolution and have become among the most prolific users of that medium for accessing, disseminating, and using data in myriad ways.[79] When data are provided as a public good via the Internet, unencumbered by proprietary rights, the positive feedback from this network externality is especially high. It becomes even greater to the extent that the data are prepared and presented online in a way that makes them available and usable to a broader range of non-expert users that extends beyond the scientific community itself.

As Stiglitz, et al., point out:

---

[76] Id., at 33.
[77] Id., at 42.
[78] See *Bits of Power*, at __.
[79] Cite statistics (?)

> The shift toward an economy in which information is central rather than peripheral may thus have fundamental implications for the appropriate role of government. In particular, the public good nature of production, along with the presence of network externalities and winner-take-all markets, may remove the automatic preference for private rather than public production. In addition, the high fixed costs and low marginal costs of producing information and the impact of network externalities are both associated with significant dangers of limited competition.[80]

These economic characteristics associated with the transmission of digital scientific data on the Internet provide a strong argument for such activities to be undertaken within the public domain by government agencies or by nongovernmental entities that receive government support. At the same time, some potentially countervailing factors that have been found to attenuate the economic efficiency and effectiveness of government actors must also be taken into account. These factors include: the lack of a bankruptcy threat, weak incentives for workers, skewed incentives for managers, the inability to make credible commitments over extended periods of time, and an aversion to risk coupled with weak incentives to innovate.[81]

The last two of these possible limitations especially illuminate the government's role in the production and dissemination of public-domain S&T data. Despite the best of intentions or the drafting of long-term plans, government managers cannot legally guarantee stable support or even continuity of any government activity beyond each current fiscal year. The budget of the federal government (and of every state) can only be legislatively appropriated for one fiscal year at a time. This makes every publicly funded activity—including the production, maintenance, and dissemination of S&T data—subject to the fiscal vagaries of the government.

---

[80] Stiglitz, et al., at 44.

[81] Id., at 35-36 and 44; citing Stiglitz (1988), at 198-212, and W.A. Niskanen (1971), *Bureaucracy and Representative Government*, Adline, Chicago.

Moreover, S&T data activities are neither entitlements, such as Social Security, Medicare, or Medicaid (which take up the largest portion of the budget and typically are the last budget category to be reduced) nor, from a political standpoint, are they high-priority discretionary budgetary items, such as public safety and defense expenses. Hence, they remain particularly vulnerable to the effects of economic downturns or of responses to national emergencies. Of course, there are no guarantees of stability or continuity in private-sector research or S&T database development either, but greater legal certainty and enforceability resides in private contracts, and even the S&T databases of bankrupt companies can be rescued for pennies on the dollar. The actual effects of these limitations on both the public and private sectors with regard to scope and management of public-domain data are addressed in more detail in Part II.

Perhaps more important, in the larger picture, is the greater inclination of government entities to risk aversion, which is reinforced by the government's greatly reduced incentives to innovate. These problems result partly from the absence of the strong motivational factor inherent in market forces and partly from the difficulties that government R&D managers encounter in making bold decisions owing to their lack of direct and long-term budgetary control, as noted above. The government's bureaucratic conservatism thus stands in stark contrast to the well-known risk taking and innovative genius of the private sector in the United States.

Tendencies like these, however, which hamper government economic activity in many spheres of the private sector, often turn out to benefit government production or provision of certain public goods and services. National defense comes immediately to mind. Basic research and the related production of scientific data appear to fall in a similar category, because their high fixed costs, low marginal costs of duplication or dissemination, and small or uncertain markets seem consonant with risk aversion and alien to the climate surrounding innovation in more applied research and technology development. The conduct of basic research is inherently risky, and sometimes even foolish, for a company that bases its investment decisions on the prospects of quick market acceptance, short-term profitability, and tangible returns to shareholders.

A lot of basic research may never yield any direct profit in the longer term. Robotic space science missions, particle physics, the study of the Earth's environment, some social and behavioral sciences, and many other data-intensive research areas often lie outside the likely, feasible scope of private market production, as measured in short-term returns. Yet these endeavors remain fundamental sources of data and information that feed our knowledge-based economy. Moreover, while most research projects conducted in the public sector do not result in major commercial payoffs, many notable advances are subsequently commercialized by the private sector and do expand the economy.[82] Some advances that flow directly from public-sector research investments, like the development of communication satellites or the Internet, end up by spawning whole new economic sectors that change the world. The shareholders—the taxpaying public—thus derive both tangible and intangible benefits, or positive network externalities, from these public research investments and their related S&T data and information products. Taken together, these economic factors favor the continued production and dissemination of basic S&T data sets, which are the raw materials of the knowledge-based economy, by government or with strong government support. This would leave both patentable and subpatentable forms of innovation, including value-adding and market-oriented applications of the government's public-domain data and information, for production and exploitation by the private sector. It is worth emphasizing in this connection that Stiglitz et al, when formulating principles for deciding in which online information activities the government should engage, propose the provision of public data and information and the support of basic research as two of only three fully supportable government functions.[83] Their conclusion is not only consonant with long-established U.S. information policies, principles, and practices concerning the public domain in government information and research; it is entirely consistent with the norms and values of most scientists who work outside the commercial private sector, as discussed above.

---

[82] For example, approximately 70 percent of patents granted in the United States in the 1990s were based, at least in part, on basic (government or government-funded) research. Narin. See also Mansfield, and OTA (1990).

[83] Stiglitz, et al. (2000), at 50-57.

II. PRESSURES ON THE TRADITIONAL FUNCTIONS OF THE PUBLIC DOMAIN

A. The Growing Commodification of Data

Economic pressures on both government and university producers of scientific data are continuing to narrow the scope of what is created or placed in the public domain, with resulting access and use restrictions on resources that were once openly available to all. The pressures on the government are both structural and political. As noted in Part I, the structure of the federal (and individual state) budgets is divided between social entitlement expenditures, such as Social Security, and other "discretionary" budget items. Because entitlements are mandated by law, are politically difficult to revise, and increase inexorably in total cost at a rate greater than the total budget, the amount spent on all other discretionary programs, including federal research, continues to shrink as a percentage of the overall budget.

This structural limitation in the federal budget is compounded by the rapidly rising costs of state-of-the art research, whether in terms of researcher salaries, scientific equipment, major facilities—or information infrastructure. With specific regard to information infrastructure, the lion's share of expenses are earmarked for computing and communications equipment, with the remainder, such as it is, devoted to managing, preserving, and disseminating the public-domain data and information resulting from basic research. The government's S&T data and information services are thus the last to be funded, and they are almost always the first to suffer cutbacks, despite the proven value those data and information products have for the research process, the economy, and society generally. For example, the NOAA's budget for its National Data Centers has remained flat and actually decreased in real dollars over the past 20 years, while its data holdings have increased exponentially and its overall budget has doubled.[84] Almost all other

[84] See BITS OF POWER, supra note __, at __.

science agencies have experienced a similar situation, with the recent exception of NASA and the NIH.[85]

These chronic budgetary shortfalls for managing and disseminating public-domain scientific data have been accompanied by recurring political pressures on the scientific agencies to privatize their outputs.[86] Until recently, the common practice of science agencies had been to procure data-collection services, such as an observational satellite or ground-based sensor system from a company, typically under a cost-plus contract and pursuant to government specifications, and frequently based on consensus requirements that the research community recommended.[87] The company would build and deliver the data collection system, which the agency would then operate, pursuant to its mission, and all the data from the system would enter the public domain.

However, the economic trends noted above, coupled with the legal trend to license rather than sell data and information (discussed in more detail in the next section), have encouraged industry to change from delivering data-collection systems to seeking to supply the government's needs for data and information products, sometimes referred to as "productization."[88] The reason is simple. Why charge one fee to deliver a technological system for the government to collect, package, and disseminate data when you can persuade the government to largely pay for that same system, but only to license the resulting data products? This solution leaves the control and ownership of those data in the hands of the company, and allows it to license them ad infinitum to anyone else willing to pay. Because of this new-found role of the government agency as cash cow, fed at the

---

[85] Add figures for some other agencies and NIH.

[86] See notes _____ *infra*, and accompanying text. In other countries, notably in the E.U., the trend has been to commercialize data right from the public source. See generally the *EC Green Paper* (1999) and Plujimers and Weiss (publication pending). Recent reports indicate that the E.U. policy may be changing toward a more open approach to its member states' government information, however

[87] [Add explanation of how NRC science strategies define research goals and supporting data specifications for various discipline areas, and cite NRC reports]

[88] MAPPS Conference, 2000.

taxpayer trough and ready to be milked, there recently has been a great deal of pressure on the science agencies, particularly through Congress, to stop collecting or disseminating data and to obtain those data from the private sector instead.

This approach has previously resulted in at least one well-documented fiasco, namely, the privatization of the NASA-NOAA Landsat program in 1985, which seriously undermined basic and applied research in environmental remote sensing in the United States for the better part of a decade.[89] More recently, the Commercial Space Act of 1998 has directed NASA to purchase space and earth science data collection and dissemination services from the private sector and to treat data as commercial commodities under federal procurement regulations.[90] Similar pressures were placed on the National Oceanic and Atmospheric Administration in Congress in the last session by the meteorological data value-adding industry,[91] and in 2001 on the Department of Energy by the Software and Information Industry Association.[92] There also are strong indications that the same type of effort will be made by the photogrammetric industry with regard to the data-collection and dissemination activities of the U.S. Geological Survey.[93] Although the full extent of these privatization initiatives is not yet known, there is reason to predict that the effects are likely to be as bad for science and other public-interest users as was true in the case of Landsat.[94]

The practice of licensing data and information products from the private sector raises serious questions about the types of controls the latter places on the redistribution and uses of such data and information that the government can subsequently undertake. When the terms of the license saddle the government with onerous obligations, and access, use, and redistribution are substantially restricted, as they almost always are, neither the

---

[89] For a discussion of the effects that the privatization of the Landsat program had on basic research, see Bits of Power (1997), at 121-124.

[90] [cites]

[91] Tallia ref

[92] Chronicle of Higher Education, July 2001

[93] MAPPS Conference, supra note __.

[94] *See supra* note _____.

agency nor the taxpayer is well served. This is particularly true in those cases where the data to be collected are meant for a basic research function or to serve a key statutory mission of the agency.[95] A similar, but no less serious problem, arises when a government agency either abdicates or outsources its data dissemination functions, which are then placed under the exclusive proprietary control of a private-sector entity. The public domain has been further reduced through the use of increasingly popular Cooperative Research and Development Agreements (CRADAs) between federal agencies and private-sector entities, in which the cooperating companies almost always retain the intellectual property rights to all research results.[96]

In the academic sector, the predominant norms remain open disclosure and the sharing of research data at the time of publication, if not before, and the placement of the data derived from federally-funded research in public data centers and archives. Nevertheless, there have been various policy incentives and steady economic pressures on research universities and academics to protect and commercialize their data, rather than to place them in the public domain. The costs of research and education activities in universities have far outpaced inflation, so there are direct economic concerns to recover costs and generate new income wherever possible.[97] Perhaps most significant, the 1980 Bayh-Dole Act has encouraged academics to protect and commercialize the fruits of their federally-funded research, especially in the potentially lucrative biomedical research area,[98] and similar laws have been passed at the state level.[99]

These pressures have led universities to adopt institutional policies and mechanisms to facilitate the creation of start-ups by faculty or of joint ventures with industry.[100] Such commercial activities are partially circumscribed by countervailing policies and formal institutional guidelines that seek to promote the educational and public-interest missions of

[95] Stiglitz, et al., supra note __, at __.
[96] CRADA regs
[97] [cite]
[98] Cite the Act. This issue is discussed in depth in the article by Eisenberg and Rai for this conference.
[99] [cites]
[100] [provide examples and cites]

universities.[101] Nevertheless, the proliferation of commercial activities in an otherwise non-commercial academic environment necessarily leads to changes in the underlying norms that foster open communication and the sharing of data and research results with faculty colleagues and students.[102] They encourage instead the proprietary protection and licensing of such data and results, and the limitation of scholarly publication.[103] Moreover, in response to the generally increasing legal protection of intellectual property, and the concomitant diminution of a clearly identifiable public domain, many universities have adopted stricter institutional rules and guidelines pertaining to access, use, and distribution of protected forms of information.

The trend in both government and academia toward the removal from the public domain of the data collection and dissemination functions that support basic research and critical government mission areas, and toward placing those data under proprietary control raises a fundamental public policy issue. This movement to commodify data suddenly shrinks the universe of potential unrestricted users from practically anyone in the world, to a small class of authorized users numbering perhaps in the dozens or hundreds.[104] If the bulk, or even a substantial fraction, of primary research data sources are shifted from the public domain to restricted proprietary sources, we will produce a situation similar to that which exists in most other countries by default, with one of the linchpins of the American system of scientific progress and innovation removed in short order.

B. The Legal Onslaught

---

[101] cite University of Maine policy, 2001

[102] See, e.g., James Robert Brown, "Privatizing the University–the New Tragedy of the Commons," *Science*, Vol. 290, 1 December 2000, pp. 1701-02.

[103] STEP cites. *See also* J. H. Reichman, Computer Programs as Applied Scientific Know-How, ____ VAND. L. REV. _____ (1989) (discussing confused university policies regarding ownership of software and other applications of unpatentable know-how to industry).

[104] *Cf.* Boyle, Second Enclosure Movement (this conference).

The digital revolution makes investors acutely aware of the heightened economic value that collections of information and data may acquire in the new information economy.[105] There were, of course, always concerns about incentives to produce basic data and information as raw materials of the innovation process, especially in light of gaps in intellectual property law that seemed to leave databases in limbo.[106] However, the dominant legal and economic paradigms focused attention primarily on downstream aggregates of information packed into sufficiently large bundles that could attract the protective monopolies of the patent and copyright laws, and the user-friendly rules of copyright laws as applied to print media did not, on the whole, unduly hinder industrial research and development.[107]

Proprietary rights in more diffuse bundles of information were largely confined to trade secret laws and general unfair competition laws, which provide liability rules against market destructive conduct that help compilers to appropriate reasonable returns from their investments.[108] These rules left most upstream flows of data unprotected by intellectual property rights and freely available as a raw material of the national system of innovation.[109]

In the new digital economy, attention has logically focused on the incentive structure for generating data and information and on the possibility that commodification of even public-sector and public-domain data would stimulate major new investments by providing new means of recovering the costs of production.[110] Moreover, investors have increasingly understood the economic potential that awaits those who capture and market data and information as raw materials or inputs into the upstream stages of the innovation process. A group of major transnational database marketers have accordingly sought stronger legal and technical means of marketing data to national innovation systems that formerly took their free availability for granted.

---

[105]*See, e.g.,* S. Maurer, Industry Canada (2001); cites.

[106]The first to spot this problem in its modern guise was Professor Robert Denicola. *See* Denicola, COLUMB. L. REV. 1980?).

[107]*See supra* text accompanying notes _____.

[108]*See* Reichman, Legal Hybrids.

[109]*See, e.g.,* Reichman & Samuelson (1997). Reichman and Uhlir (1999).

[110]*See, e.g.,* Tyson.

1. *Sui Generis* Intellectual Property Rights Restricting the Availability of Data as Such

When it came to formulating a regulatory regime for noncopyrightable databases, the Commission of the European Communities embarked down an entirely new path. Apparently, it saw – or thought it saw – an opportunity to jump start an important industry whose participants in Continental countries had lagged behind their counterparts in the English-speaking countries, particularly the U.S. and the U.K. The initial problem of how to harmonize different approaches to a perceived gap in the law, which may have left some database producers vulnerable to free riders, thus seems to have given way over time to a regulatory design that aimed, at least in good measure, to expand the share of European producers in the growing global market for databases at the expense of producers in other countries.[111] In so doing, the Commission gradually gave birth to a new and unprecedented form of intellectual property protection that exceeds the protectionist boundaries that have heretofore limited either the dominant patent and copyright paradigms or the deviant hybrid regimes of exclusive property rights taken as a class.

a. The E.U. Database Directive in brief

The *sui generis* regime that the Commission ultimately adopted in its Directive on the Legal Protection of Databases in 1996[112] is like nothing we have ever seen before. It protects any collection of data, information, or other materials that are arranged in a systematic or methodological way, provided that they are individually accessible by electronic or other means. This does not, however, imply that some organized form of storage is needed.[113] The criterion of eligibility is a "substantial investment," as measured in either qualitative or quantitative terms, and the courts are left to develop this concept.[114] That the drafters believe a relatively minimal level of investment would suffice appears from

---

[111] *See* Steven Maurer (2001).

[112] Directive 96/9/EC of the European Parliament and the Council of 11 March 1996 on the Legal Protection of Databases 1996 O.J. (L77) 20 [hereinafter EC Directive].

[113] Hugenholz (2000).

[114]For subsequent developments in the early cases, *see* Hugenholz (2000).

an explicit recognition that the qualifying investment may consist simply of verifying or maintaining the database.[115]

In return for this investment, the compiler obtains exclusive rights to extract or to utilize all or a substantial part of the contents of the protected database. The exclusive extraction right pertains to any transfer in any form of all or a substantial part of the contents of a protected database; the exclusive reutilization right covers only the making available to the public of all or a substantial part of the same database.[116] In every case, the first comer obtains an exclusive right to control uses of raw data as such, as well as a powerful adaptation (or derivative work) right along the lines that copyright law bestows on "original works of authorship," even though such a right is alien to the protection of investment under existing unfair competition laws.[117]

The Directive provides no major public-interest exceptions, comparable to those recognized under domestic and international copyright laws. An optional, but ambiguous, exception concerning illustrations for teaching or scientific research is said to be open to flexible interpretation, and some member countries have implemented it in different ways. Other countries have simply ignored this exception altogether, which contradicts the Commission's supposed concerns about uniform law.[118] Moreover, European governments that generate data may exercise either copyrights or *sui generis* rights in their own productions. This contrasts with the situation in the United States, where the government cannot claim intellectual property rights in the data it generates and must make such data available to the public for no more than a cost- of-delivery fee.[119]

The Directive's *sui generis* regime does exempt from liability anyone who extracts or uses an insubstantial part of a protected database. However, such a user bears the risk

---

[115][cites]

[116]*See* Hugenholz (2000); Maurer (2001)

[117]*See* Reichman & Samuelson (1997).

[118][cites] One should note that one of the principal lobbyists supporting strong database protection in both the E.U. and the U.S. is the world's leading supplier of commercialized scientific publications.

[119]*See* NRC, A QUESTION OF BALANCE (199); Reichman & Uhlir (1999); *supra* notes _____ and accompanying text.

of accurately drawing the line between a substantial and an insubstantial part,[120] and any repeated or systematic use of even an insubstantial part will forfeit this exemption.

Qualifying databases are nominally protected for a fifteen-year period. In reality, each new investment in a protected database, such as the provision of updates, will requalify that database as a whole for a new term of protection. In this and other respects, the *sui generis* adaptation right is far more powerful than that of copyright law, which attaches only to the new matter added to an underlying, pre-existing work and limits the term of that protection.

Finally, the Directive carries no national treatment requirement into its *sui generis* component. Foreign database producers become eligible only if their countries of origin provide a similar form of protection or if, in keeping with a goal attributed to the Commission, they set up operations within the E.U.[121]

The E.U.'s Directive on the Legal Protection of Databases thus broke radically with the historical limits of intellectual property protection in at least three ways:

1. It overtly and expressly conferred an exclusive property right on the fruits of investment as such, without predicating the grant of protection on any pre-determined level of creative contribution to the public domain;
2. It conferred this new exclusive property right on aggregates of information as such, which had heretofore been considered an unprotectible raw material or basic input available to creators operating under all other pre-existing intellectual property rights;
3. It potentially conferred the new exclusive property right in perpetuity, with no concomitant requirement that the public ultimately acquire ownership of the object of protection at the end of a specified period.

---

[120] *See* Reichman & Samuelson (1997).

[121] *See, e.g.*, Maurer (2001). However, nonqualifying foreign producers may continue to invoke the residual domestic copyright and unfair competition laws, where available, and the cases so far arising under the various members' implementing statutes suggest that both regimes may remain available to foreign parties. *See, e.g.* Hugenholz (2000). A detailed discussion of the various implementing statutes and of the case law to date is beyond the scope of this paper.

In this and other respects, the E.U.'s Database Directive broke with the history of intellectual property law by allowing a property rule – as distinct from a liability rule – to last in perpetuity and by abolishing the very concept of a public domain that had historically justified the grant of temporary exclusive rights in intangible creations.[122]

b. The database protection controversy in the United States

The situation in the United States differs markedly from that which preceded the adoption of the European Commission's Directive on the Legal Protection of Databases. In general, the legislative process in the U.S. has become relatively transparent. Since the first legislative proposal, modeled on the E.U. Directive, was introduced by the House Committee on the Judiciary in May 1996,[123] this transparency has generated a spirited and often high-level public debate.[124]

The resulting controversy has, in turn, led to the crystallization of two opposing coalitions that favor very different approaches.[125] Although forced negotiations among the stakeholders have been underway since April 2001, and the principal committee chairmen have vowed to draft a compromise bill if the interested parties themselves fail to agree, very little progress toward a compromise solution had been reached as of the time of writing. Given the intensity of the opposing views, the methodological distance that divides them, and the political clout of the opposing camps, this is hardly surprising. Whether some breakthrough will eventually occur cannot be safely predicted here, nor is there any credible basis for predicting the shape such a breakthrough might assume were it to occur.

We are, accordingly, left with the two basic proposals that were still on the table at the end of the last legislative session, which ended in an impasse. These proposals, as refined during that session, represent the baseline positions that each coalition carried into the current round of negotiations. One bill, H.R. 354, as revised in January, 2000,

---

[122] Reichman & Samuelson (1997).

[123] H.R. 3534. For details, see Reichman & Samuelson (1997).

[124] See, generally, Reichman & Uhlir (1999)

[125] *See* J. H. Reichman, Database Protection in the Global Economy (2001). For developments in the period 1997-1999, *see* Reichman and Uhlir (1999).

embodies the proponents' last set of proposals for a *sui generis* regime built on an exclusive property rights model (although some effort has been made to conceal that solution behind a facade that evokes unfair competition law). The other bill, H.R. 1858, sets out the opponents' views of a minimalist misappropriation regime as it stood on the eve of the current round of negotiations.

(i) The exclusive rights model

The proponents' exclusive property rights model embodied in H.R. 354 attempts to achieve levels of protection comparable to those of the E.U. Directive by means that are somewhat more congenial to the legal traditions of the United States. The changes introduced at the end of the last legislative session, in particular softened (often under pressure from representatives of the previous Administration seeking to engender a compromise) some of the most controversial provisions at the margins, while maintaining the overall integrity of a strongly protectionist regime.[126] Despite further concessions that

[126]The bill in this form continues to define "collections of information" as "a large number of discrete items of information ... collected and ... organized for the purpose of bringing discrete items of information together in one place or through one source so that persons may access them." (§1401(1)). This definition is so broad, and the overlap with copyright law so palpable, that it is hard to conceive of any assemblage of words, numbers, facts or information that would not qualify as a potentially protectable collection of information.

Like the E.U. Directive, this Bill casts eligibility in terms of an "investment of substantial monetary or other resources" in the gathering, organizing or maintaining of a "collection of information." (§1402(a)). It then confers two exclusive rights on the investor, viz., a right to make all or a substantial part of a protected collection "available to others" and a right "to extract all or a substantial part to make available to others." Here the terms "others is manifestly broader than "public" in ways that remain to be clarified. However, the second right represents a concession to the past Administration in that it foregoes the general right to control private use that appeared in previous versions. This concession thus reduces the scope of protection to a point more in line with the E.U.'s reutilization right, and it does not impede personal use by one who lawfully acquires access to the database. (§1402(a)).

H.R. 354 then superimposes an additional criterion of liability on both exclusive rights that is not present in the E.U. model. This is the requirement that, to be liable, any unauthorized act of "making available to others" or of "extraction" for that purpose must cause "material harm to the market" of the qualifying investor "for a product or service that incorporates that collection of information and is offered or intended to be offered in commerce."(§1402(a)). The crux of liability under the Bill thus derives from a "material harm to markets" test that is meant to cloud the copyright-like nature of the Bill and to shroud it in different terminology. In fact, a "harm to markets" test is lifted

were made to the opponents' concerns in the last iteration of the bill (January 11, 2000), some of them real, others nominal in effect, the bill effectively ensures that first comers will control the extraction and distribution of raw data as such, as well as follow-on applications-- "derivative works" --in virtually all cases.

The bill introduces a "reasonable use" exception that, in one sentence seems to benefit the non-profit user communities, especially researchers and libraries, and that is meant to convey a sense of similarity with the "fair use exception" in copyright law.[127] In reality, virtually every customary or traditional use of facts or information compiled by others that copyright law would presumably have allowed scientists, researchers, or other nonprofit entities to make in the past now become prima facie instances of infringement under H.R. 354. These users would in effect either have to license such uses or be prepared to seek judicial relief for "reasonableness" on a continuing basis. Because universities dislike litigation and are risk averse by nature, and
this provision puts the burden of showing reasonableness on them, there is reason to expect a chilling effect on customary uses of data by these institutions in te event that such a bill is eventually enacted.[128]

As more and more segments of industry come to appreciate the market power that major database producers could thus acquire under the proposed legislation, one after another has petitioned the subcommittee for special relief. Thus, this bill, which has now grown to some thirty pages in length, singles out various special interests who benefit, to varying degrees, from special exemptions from liability.[129] Government-generated data

---

bodily from §107(4) of the Copyright Act of 1976, and it reflects the better view of what U.S. copyright law is all about. See Reichman, Goldstein on Copyrights.

[127] H.R. 354, §1403.

[128] A further provision then completes the sense of circularity by expressly exempting any nonprofit educational, scientific, and research use that "does not materially harm the market" as previously defined (§1403(b)). Since any use that does not materially harm the market remains unactionable to begin with, this "concession" adds nothing but window dressing. However, another vaguely worded exception seems to recognize at least a possibility that certain "fully transformative uses" might nonetheless escape liability, but this ambiguous exception defies interpretation in its present form and remains to be clarified.

[129] At the time of writing the list of those entitled to such immunities included news reporting organizations (1403(3)); churches that depend on genealogical information, notably the Mormons (1403(i)); online service providers; certain stockbrokers; and to a still unknown extent, nonprofit

remain excluded, in principle, from protection, in keeping with current U.S. practice.[130] However, there is considerable controversy concerning the degree of protection to be afforded government-generated data that subsequently become embodied in value-adding, privately funded databases.[131]

The bill imposes no restrictions whatsoever on licensing agreements, including agreements that might overrule the few exceptions otherwise allowed. Despite constant remonstrations from opponents about the need to regulate licensing in a variety of circumstances, and especially with respect to sole-source providers,[132] the bill itself has not budged in this direction. On the contrary, new provisions added to the last iteration of H.R. 354 would set up measures prohibiting tampering with encryption devices ("anti-circumvention measures") and with electronically embedded or "watermarked" rights management information, in a manner that parallels the provisions adopted for online transmissions of copyrighted works under the Digital Millennium Copyright Act of 1998.[133] Because these provisions effectively secure the database against unauthorized access (and tend to create an additional "exclusive right to access" without expressly so declaring), they would only add to the database owner's market power to dictate contractual terms and conditions without regard to the public interest.[134]

The one major concession that has so far been made to the opponents' constitutional arguments concerning the question of duration. As previously noted, the E.U. Directive allows for perpetual protection of the whole database so long as any substantial part of it is updated or maintained by virtue of a new and substantial investment,

---

research organizations.

[130] H.R. 354, §1404.

[131] All parties agree that a private, value-adding compiler should obtain whatever degree of protection is elsewhere provided, notwithstanding the incorporation of government-generated data. The issue concerns the rights and abilities of third parties to continue to access the original, government-generated data sets, notwithstanding the existence of a commodified embodiment. At the time of writing, the proponents were little inclined to accept measures seeking to preserve access to the original data sets, but pressures in this direction were building.

[132] *See, e.g.*, Reichman & Uhlir (citing authorities).

[133] [cites]

[134] These powers are further magnified by the imposition of strong criminal sanctions in addition to strong civil remedies for infringement, which can run concurrently with any additional penalties for copyright infringement that may be awarded to a plaintiff.

and the proponents' early proposals in the U.S. echoed this provision.[135] However, the U.S. Constitution clearly prescribes a limited term of duration for intellectual property rights, and the proponents have finally bowed to pressures from many directions by limiting the term of duration to fifteen years. Any update to an existing database would then qualify for a new term of fifteen years, but this protection would apply, at least in principle, only to the new matter added in the update.[136]

(ii) The misappropriation model

The opponents' own bill, H.R. 1858,[137] was first put before the House Commerce Committee in May 1999, and significantly amended later in the year, as a sign of good faith. The underlying purpose of this bill was to prohibit wholesale duplication of a database as a form of unfair competition. It thus set out to create a minimalist liability rule that prohibits market-destructive conduct rather than an exclusive property right as such, and in this sense, it posed a strong contrast to H.R. 354.[138] A later iteration of the bill, designed to win supporters away from H.R. 354, made H.R. 1858 surprisingly protectionist in possibly unintended ways, as will be seen below. Moreover, the realities of the bargaining process are such that concessions unwisely made to the high protectionist camp at an earlier stage, for whatever tactical reasons, are unlikely to be able to be withdrawn now.

---

[135] *See supra* note _____ and accompanying text; Reichman & Samuelson (1997).

[136] H.R. 354, §1409(i). In practice, however, the inability to clearly separate old from new matter in complex databases, coupled with ambiguous language concerning the scope of protection against harm to "likely, expected, or planned" market segments may still leave some loophole for an indefinite term of duration.

[137] [cites. First offered May 19, 1999 but significantly amended in July of that year].

[138] H.R. 1858 begins with a definition of databases that is not appreciably narrower than that of H.R. 354, except for an express exclusion of traditional literary works that "tell a story, communicate a message," and the like (§101(1)). In other words, there is at least some attempt to draw a clearer line of demarcation between the proposed database regime and copyright law, and to reduce overlap or cumulative protection as might occur under H.R. 354.

The operative protective language in H.R. 1858 appears short and direct, but it relies on a series of contingent definitions that muddy the true scope of protection. Thus, the Bill would prohibit anyone from selling or distributing to the public a database that is 1) "a duplicate of another database ... collected and organized by another person or entity" and 2) is sold or distributed in commerce in competition with that other database." (§102). A prohibited duplicate is then defined as a database that is "substantially the same as such other database, as a result of the extraction of information from such other database." (§101(2))

Liability under H.R. 1858 attaches in the first instance only for a wholesale duplication of a pre-existing database that results in a substantially identical end product. However, this basic misappropriation approach becomes further subject to both expansionist and limiting thrusts. Expanding the potential for liability is a proviso added to the definition of a protectable database that treats "any discrete sections [of a protected database] containing a large number of discrete items of information" as a separably identifiable database entitled to protection in its own right.[139] The bill would thus codify a surprisingly broad prohibition of follow-on applications that make use of discrete segments of pre-existing databases, subject to the limitations set out below.

A second protectionist thrust results from the lack of any duration clause whatsoever. In other words, the prohibition against wholesale duplication – subject to limitations set out below – could conceivably last forever. This perpetual threat of liability would attach to wholesale duplication of even a discrete segment of a pre-existing database, if the other criteria for liability were also met. However, liability for wholesale duplication of all or a discrete segment of a protected database does not attach unless the unauthorized copy is sold or distributed in commerce and "in competition with" the protected database.[140] Hence, even a wholesale duplication that did not substantially decrease expected revenues (as might occur from, say, nonprofit research activities) or that did not significantly impede the investor's opportunity to recover his or her initial investment (as might occur in the case of a follow-on product sold in a distant market segment that required a substantial independent investment) could both presumably escape liability in appropriate circumstances.

The bill then further reduces the potential scope of liability by imposing a set of well-defined exceptions and also by limiting enforcement to actions brought by the Federal Trade Commission.[141] An additional set of safeguards emerges from the drafters' real

---

[139] [cite]

[140] The term "in competition with," when used in connection with a sale or distribution to the public, is then defined to mean that the unauthorized copy "substantially decreases the revenue" that the first comer otherwise expected and "significantly threatens ... [his or her] opportunity to recover a reasonable return on the investment" in the duplicated database. Both prongs must be met before liability will attach.

[141] [cites]

concerns about potential misuses of even this so-called minimalist form of protection. These concerns are expressed in a provision that expressly denies liability in any case where the protected party "misuses the protection" that H.R. 1858 affords[142]. A second provision on this topic then elaborates a long and detailed list of criteria that courts could use as guidelines in particular cases in order to determine whether an instance of misuse had occurred.[143] These guidelines or standards would greatly clarify the line between acceptable and unacceptable licensing conditions, and if enacted, they could make a handsome contribution to the doctrine of misuse as applied to other intellectual property rights.

c. *International implications*

The international implications of the various database proposals are largely beyond the scope of this paper. In general, two major possibilities are foreseen. One is that the E.U and the U.S. could align their database protection regimes if the U.S. adopted a high protectionist proposal for a strong exclusive property rights along the lines set out in H.R. 354. In that event, there would be a risk of premature harmonization in the rest of the world, with developing countries left to shoulder the resistance.

If, instead, the U.S. adopted a softer misappropriation approach, as in H.R. 1858, then there would be some possibilities of a database war between adherents of U.S. and E.U. approaches. In this connection, the E.U. requires all affiliated countries and would-be affiliates to adopt its *sui generis* database regime, a total of some fifty countries, and it seeks to impose this model in bilateral and regional trade agreements. However, these tensions could be alleviated by an umbrella treaty with a menu of options, for which there is an historical precedent.[144] This solution would require a minimalist consensus against wholesale duplication of databases, together with a transitional period of experimentation in which different states proceeded to develop their own approach.[145]

---

[142] [cites]
[143] [cite and quote in full]
[144] *Cf.* Geneva Phonograms Treaty of 1975.
[145] For details, *see* Reichman, Database Protection in a Global Economy (2001).

Under any of these solutions a period of considerable tension is likely to hamper international exchanges of data in coming years. There are abundant signs that scientific exchanges will also be effected in part because governments in E.U. countries can directly protect and exploit their data and in part because more and more scientific and technical data will be commercialized under any of these proposals.[146]

2. Changes in Existing Laws that Reduce the Public Domain in Data

While proposals to confer strong exclusive property rights on noncopyrightable collections of data constitute the clearest and most overt assault on the concept of a public domain that has fueled both scientific endeavors and technological innovation in the past, other legal developments, taken singly or collectively, could prove no less disruptive. For limitations of space, we will briefly note the impact of selected developments in both federal statutory copyright law and in contract laws at the state level that we deem most worthy of attention.[147]

a. Expanding copyright protection of compilations: the revolt against Feist

The quest for a new legal regime to protect databases was triggered in part by the U.S. Supreme Court's 1991 decision in Feist Publications, Inc. v. Rural Telephone Service Co.,[148] which denied copyright protection to the white pages of a telephone directory. That decision was notable for defending third-party access to data in two ways. At the eligibility stage of an action for copyright infringement, the Court required a compiler to show that his or her selection or arrangement of contents amounted to an original work of authorship.[149] Equally important, the court denied protection of the compiler's disparate facts at the infringement stage,[150] and limited the scope of copyright protection to the original elements of selection and arrangement that met the test of eligibility. In effect, this meant that second comers who developed their own criteria of selection and arrangement

---

[146] *Feist* supra, note __..
[147] For a more complete list of some twenty-three legal, economic, and technological assaults on the public domain, *see* Reichman & Uhlir, Assaults on the Public Domain (unpublished, 2000)
[148] [cite].
[149] 17 U.S.C. §§ 102(a), 103.
[150] 17 U.S.C. § 102(b), 103.

could in principle use prior data to make follow-on products without falling afoul of the copyright owner's strong exclusive right to prepare derivative works.[151]

In recent years, however, judicial concerns about the compilers' inability to appropriate the returns from their investments have led leading federal appellate courts to broaden copyright protection of low authorship compilations[152] in ways that significantly deform both the spirit and the letter of *Feist*. At the eligibility stage, so little in the way of originality is now required that the only print media compilations still certain to be excluded are the white pages of telephone directories.[153] More tellingly, the courts have increasingly perceived the eligibility criteria of selection and arrangement as pervading the data themselves, in order to restrain second comers from using preexisting data sets to perform operations that are functionally equivalent to those of an initial compiler. In the Second Circuit, for example, a competitor could not assess used car values by the same technical means as those embodied in a first-comer's copyrightable compilation, even if those means turned out to be particularly efficient.[154] Similarly, the Ninth Circuit prevented even the use of a small amount of data from copyrighted compilation that was essential to achieving a functional result.[155]

Opponents of *sui generis* database protection in the United States cite these and other cases as evidence that no *sui generis* database protection law is needed.[156] In reality, these cases suggest that, in the absence of a suitable minimalist regime that could cure the risk of market failure without impoverishing the public domain, courts tend to convert copyright law into a roving unfair competition law that can protect algorithms and other functional matter for very long periods of time and that could create formidable barriers to entry. This tendency, however, ignores the historical limits of copyright protection and undermines the border with patent law, in defiance of well-established Supreme Court precedents.[157]

---

[151] 17 U.S.C. §§ 101, 103, 106(2).
[152] *See* Jane Ginsberg (I); *see also* Jane Ginsberg (II).
[153] [cites].
[154] CCC v. MacClean. *But see*, Baker v. Selden, ____ U.S. _____ (1879).
[155] [cite]; *see generally*, Justin Hughes, Creating Facts, (2001).
[156] [Submissions to negotiations].
[157] Baker v. Selden; Reichman (1989), at n. 188 (deeper meaning of Baker v. Selden).

b. The DMCA: An exclusive right to access copyrightable compilations of data?

As noted in Part I, traditional copyright law was friendly to science, education and innovation by dint of its refusal to protect either facts or ideas as eligible subject matter; by limiting the scope of protection for compilations and other factual works to the stylistic expression of facts and ideas; by carving out express exceptions and immunities for teaching, research and libraries; and by recognizing a catch all, fall-back "fair use" exception for nonprofit research and other endeavors that advanced the public interest in the diffusion of facts and ideas at relatively little expense to authors. These policies were reinforced by judge-made and partially codified exceptions for functionally dictated components of literary works, which take the form of non-protectable methods, principles, processes, discoveries, and the like.[158] As we have seen, however, recent judicial decisions have cut back on this tradition even as regards compilations of data disseminated in hard copies.

With respect to copyrightable compilations of data distributed online, moreover, amendments to the Copyright Act of 1976, known as the Digital Millennium Copyright Act of 1998,[159] seem to have greatly reduced these traditional safeguards by instituting a *de facto* exclusive right of access that appears immunized from many of these traditional defenses.[160] In effect, the DMCA allows copyright owners to surround their collections of data with technological fences and electronic identity marks buttressed by encryption and other, digital controls that force would-be users to enter the system through an electronic gateway.[161] In order to pass through the gateway, users must accede to electronic contracts of adhesion, which impose the copyright owner's terms and conditions without regard to the traditional defenses and statutory immunities of the copyright law. Attempts to bypass these electronic barriers in the name of pre-existing legal defenses constitute an independent basis of infringement,[162] which does not necessarily give rise to even the most well-established traditional defenses, including the idea-expression defense and fair use.[163]

---

[158] *See* Baker. V. Selden and progeny; *supra* note _____.
[159] [cites, § 1201].
[160] *See, e.g.*, Napster. *See generally*, DIGITAL DILEMMA; Samuelson [this conference]; Litman (2001).
[161] DIGITAL DILEMMA; Julie Cohen, Lochner Revisited.
[162] [cite]
[163]

It is too soon to know how far owners of copyrightable compilations can push this so-called right of access[164] at the expense of research, competition, and free speech without incurring resistance sounding in the misuse doctrine of copyright law, the public policy and unconscionability doctrines of states contract laws, and in first amendment concerns that have in the past limited copyright protection of factual works.[165] For the foreseeable future, nonetheless, the DMCA empowers owners of copyrightable collections of facts to contractually limit online access to the pre-existing public domain in ways that contrast drastically with the traditional availability of factual contents in printed works.

c. Online delivery of noncopyrightable collections of data: privately legislated intellectual property rights.

Proprietors of copyrightable compilations of data who invoke the online advantages of the DMCA must presumably still meet the originality requirement of copyright law. This requirement, never more than modest even under *Feist*, has become still more porous in recent cases, as demonstrated above. It could nonetheless suffice to bar many databases of particular scientific or technical interest from protection in copyright law on the grounds that they partake of random and complete assortments of data that lack any creativity or "original" criteria of selection or arrangement.[166] In such cases, the ability of proprietors to emulate the DMCA by surrounding their noncopyrightable collections of data with electronic fences and other technical protection measures depends on state contract laws, as reinforced by federal laws that prohibit electronic theft in general.[167]

Once again, the purpose of the electronic fence or encryption device is to force the user through an electronic gateway, at which point he or she gains access to the noncopyrightable database only by acquiescing to the terms and conditions of a "click on" adhesion contract.[168] To the extent that these contracts, which are good against the world

---

[164] Cite article, What Right of Access?

[165] Lemley (CALIF. L. REV.); Samuelson (CALIF. SYMPOSIUM); Reichman & Franklin.

[166] *See* Reichman & Uhlir (citing authorities).

[167] [cites]

[168] The same effect is achieved by a so-called "shrink wrap" license that runs with a product or information technology, such as a computer program or a CD-ROM. *See, e.g.,* Vault v. Quaid; Pro-CD v. Zeidenberg. *See also* Management Jane Radin [licenses that run with goods].

(at least on a one-by-one basis), are allowed to impose terms and conditions that ignore the goals and policies of the federal intellectual property system, they establish privately legislated intellectual property rights that are unencumbered by concessions to the public interest.[169] For example, they will forbid all unauthorized uses, including follow-on applications or equivalents of reverse engineering, even when such uses might be permitted by federal copyright laws or by state trade secret laws.[170] By the same token, a privately generated database protected by technical devices and adhesion contracts is subject to no state-imposed duration clause, and it will, accordingly, never lapse into the public domain.

The validity of "click on" and "shrink wrap" adhesion contracts as enforceable contracts has been an open question for many years, especially with regard to sales of computer software and, lately, electronic databases as well.[171] The most recent line of cases, led by the Seventh Circuit's opinion in *Pro-CD v. Zeidenberg*,[172] has tended to validate such contracts. This line of cases brushes aside both the technical obstacles to formation in general contracts law and arguments invoking conflicts with federal intellectual property policies that would seem to trigger either the public policy defense in contracts law or the doctrine of pre-emption. In this regard, the Uniform Law Commissioners have proposed a Uniform Computerized Information Transactions Act (UCITA), which would broadly validate electronic contracts of adhesion and largely immunize them from legal challenge.[173]

If present trends continue unabated, the prospects are that privately generated information products that are delivered online -- including databases and computer software -- can be kept under a kind of perpetual, mass-market trade-secret protection, subject to no reverse engineering efforts or public-interest uses that are not expressly sanctioned by contractual licensing agreements.[174] Contractual rights of this kind, backed by a totally one-sided regulatory framework, such as UCITA, could conceivably produce an even higher level of protection than available from some future federal database right subject to statutory public-interest exceptions. The most powerful proprietary cocktail

---

[169] *See generally*, Reichman & Franklin.
[170] *Id*., at _____.
[171] *See generally* McManis; Lemley; Samuelson [CALIF. SYMPOSIUM].
[172] [cites]
[173] [cites; CALIF. SYMPOSIUM, Parts I & II.]
[174] *See, e.g*,. NRC, DIGITAL DILEMMA.

of all would probably emerge from a combination of a federal database right with UCITA-backed contracts of adhesion.

Under such a regime, which would effectively impede second comers from competing on the basis of follow-on, value-adding applications of existing databases, new entrants could enter the market only by generating new data and recreating existing databases from scratch. The available evidence suggests how difficult this can be even in general markets for nonscientific information products.[175] On the whole, markets for databases have exhibited a niche-like character, which tends to make the possibilities of recuperating the costs of generating a second entire database from scratch in order to compete with an existing database appear either physically impossible or extremely risky. At the same time, the ability of owners of existing complex databases to update and integrate them at lower costs than would-be competitors constitutes a comparative advantage that progressively bars entry even to such well-heeled competition.[176] For this and other reasons, the sole-source structure of the database industry was a characteristic that worried even the high-protectionist Commission of the European Communities.[177]

This tendency to niche markets and sole-source producers is very pronounced in the market for scientific and technical databases, whether the public or private sectors are at issue. In most cases, complex databases of interest to science cannot, either as a physical observational reality or as an economic reality be regenerated from scratch on any viable basis.[178] The scientific tradition is not to foster such duplication, but to encourage the sharing of data and the construction of new databases to address ever-deeper layers of research questions from multiple, existing databases.[179]

---

[175] [cites re legal databases, for example; Maurer].

[176] Benkler [on databases].

[177] *See, e.g.* Justin Hughes, Creating Facts, 2001.

[178] Reichman & Uhlir (1999). Sometimes the database cannot be reconstituted because the underlying phenomenon are one-time events. At other times, key components of a complex database cannot feasibly by regenerated at a later date or under other conditions.

[179] This tradition of access to upstream data underlies both the existing scientific structure and the national system of innovation. Needless to say, this tradition (and the practices it supports), is threatened by the growing tendency to privatize data by the means described in this section. *See* Reichman and Franklin.

How to regulate privately generated databases made available to the public by online delivery and electronic contracts of adhesion will become a serious question no matter what federal database legislation is enacted and no matter what the fate of UCITA becomes. This problem will require courts and legislatures to develop new concepts of "misuse of contractual rights" to bridge the gap between private and public interests, and especially to promote competition, research, and free speech.[180] For present purposes, the likelihood that more and more databases of importance to science and technological development are likely to become privatized and made available only in encrypted formats with onerous restrictions on use constitutes further pressing evidence that science must take steps to organize its own means of accessing and distributing data, regardless of legislative and legal developments in other spheres of activity.

C. Technological Straightjackets and Memory Holes

As discussed in greater detail in other presentations at this Conference, highly restrictive digital rights management technologies are being developed, such as hardware and software based "trusted systems," online database access controls, digital watermarks, and increasingly effective forms of encryption.[181] These emerging technological controls on content, when combined with the changes to the intellectual property laws and institutional practices noted above, can completely supersede long-established user rights and exceptions under copyright law for print media, thereby eliminating large categories of data and information from public-domain access.[182] This is especially insidious in the context of scientific research, most of which relies on open access and liberal uses of scientific data for advancement.

In addition, there are inherent weaknesses of digital technologies with regard to the long-term preservation of data and information, including the fairly rapid deterioration of storage media, frequent changes in commercial media standards, and format incompatibilities.[183] When combined with poor information management practices, lack of resources for preservation, and ever-longer periods of proprietary protection, these factors

---

[180] *See* Reichman & Franklin (proposed a "public interest unconscionability doctrine" for contracts of adhesion on state contracts law).

[181] See Stefik (1999), NRC (1999).

[182] Lessig.

[183] See NRC 1995, plus other newer cites

together could lead to the deterioration or complete loss of large amounts of digital data and information. This is a problem both for proprietary databases before their statutory periods of protection lapse (assuming that they are not guarded in perpetuity by contracts and technological barriers), as well as for databases originally created in or transferred to the public domain.

In the case of proprietary materials, both the licensing practices and the increasingly excessive periods of protection can result in their ultimate loss, absent a strong commitment by the rights holder to properly preserve that material or to place archival copies in multiple public repositories. This problem is particularly acute for databases, many of which are continuously updated and dynamic. Whereas there is a well-established archival deposit procedure for copyrighted works, despite its erosion through the private licensing of increasing numbers of such works, the situation with regard to the similar deposit of proprietary databases, whether copyrighted or not, is much less settled. Since many databases are continually changing, no official archival copy may be said to exist for deposit. Moreover, by far the most practical and useful way to make proprietary dynamic databases available is in digital form online, for which licensing is the optimal means of protection. This further reduces the likelihood that an archival copy will ever be deposited, because there may be little incentive for the rights holder to preserve such a database in the public domain once the activity is terminated or the rights holder goes out of business. One potentially promising solution to help guarantee the indefinite preservation of proprietary databases is to establish a trust fund for such a purpose, financed by a small tax by the vendor on the user fees.[184]

However, even for databases that are created or collected by government or through government funding, there is no guarantee of preservation, and there are many instances already in which large and irreplaceable data sets have been lost. For example, data from many of the early space science and Earth observation missions conducted by NASA are gone, as are the data from the initial meteorological satellites operated by NOAA,[185] although these problems have generally been rectified in recent years.

---

[184] The American Geophysical Union is reportedly establishing such a fund for the preservation of its electronic journals (personal communication from Fred Spilhouse, 2001),

[185] General Accounting office (1989, 1991). It should be noted, however, that both NASA and NOAA, much to their credit, have since taken strong measures to rectify these earlier data preservation problems. Indeed, perhaps the greatest threat to the public domain in digital federal government records is the National Archives and Records Administration itself, which has a

D. Impact of a Shrinking Public Domain

The foregoing analysis has documented a broad range of economic, legal, and technical pressures on the continued availability of data in a public domain accessible to all users. The point of the present section is to emphasize how radical a change we are about to make in our national system of innovation, and to consider how great the risks of such a change really are.

1. A Market-Breaking Approach

As described in Part I, the U.S. system of innovation is largely premised on enormous flows of mostly government-generated or government-financed scientific and technical data, which everyone is free to use, and on free competition with respect to downstream information goods. Traditionally, United States intellectual property law did not protect investment as such; and it did not protect even privately generated upstream flows of information that were publicly distributed in hard copies except by copyright law (with its public-interest exceptions) or by the liability rules of trade secret law or general unfair competition law (which permit reverse engineering by honest means and follow-on applications that are the fruit of independent investments).[186]

The classical intellectual property system protected downstream bundles of information in two situations only: copyrightable works of art and literature, and patentable inventions. However, the following conditions apply:

- These regimes both require relatively large creative contributions based on otherwise free inputs of information and ideas;
- They both presuppose a flow of unprotected information and data upstream;
- They both presuppose free competition as to the products of mere investment that are neither copyrightable nor patentable[187].

---

minuscule budget (less than $5M/year) and a small staff for the permanent preservation of all the nation's federal records!

[186]*See supra* note _____.

[187]Sears-Compco (1964); Bonito Boats (1989).

As previously observed, the E.U.'s Database Directive changes this approach, as would a pending parallel proposal to enact *sui generis* database rights in the U.S. that is now before Congress. Specifically, the *sui generis* database regimes confer a stronger and, in the E.U., potentially perpetual exclusive property right in the fruits of mere investment, without requiring any creative contribution; and they convert data and technical information as such, which are the raw materials or basic inputs of the modern information economy and which were previously unprotectable, into the subject matter of this new exclusive property right.

The *sui generis* database regimes would thus effectuate a radical change in the economic nature and role of intellectual property rights (IPRs). Until now, the economic function of IPRs was to make markets possible where previously there existed a risk of market failure due to the public-good nature of intangible creations. Exclusive rights make embodiments of intangible public goods artificially appropriable, they create markets for those embodiments, and they make it possible to exchange payment for access to these creations.[188]

In contrast, an exclusive intellectual property right in the contents of databases breaks existing markets for downstream aggregates of information, which were formed around inputs of information largely available from the public domain. It conditions the very existence of all traditional markets for intellectual goods on:

- the willingness of information suppliers to supply at all (they can hold out or refuse to deal),
- the willingness not to charge excessive or monopoly prices (i.e., more than downstream aggregators can afford to pay in view of their own risk management assessment), and
- the willingness and ability of information suppliers to pool their respective chunks of information in contractually constructed cooperative ventures.

This last complication is perhaps the most telling of all. In effect, the *sui generis* database regimes create new and potentially serious barriers to entry to all existing markets for intellectual goods owing to the multiplicity of new owners of upstream information in whom they invest exclusive rights, any one of whom can hold

[188] *See supra* notes _____ and accompanying text.

out and all of whom can impose onerous transaction costs (analogous to the problem of multi-media transactions under copyright law). This tangle of rights is known as an anti-commons effect, and the database laws appear to be ideal generators of this phenomenon.[189]

There is, in short, a new built-in risk that too many owners of information inputs will impose too many costs and conditions on all the information processes we take for granted in the information economy. At best, the costs of research and development activities seem likely to rise across the entire economy, well in excess of benefits, owing to the potential stranglehold of data suppliers on raw materials. This stranglehold will increase with market power as most databases are owned by sole-source providers (especially in science and technology).[190]

Incurring these risks of disrupting or deforming the national system of innovation is
hardly justified by the potential social gains of a strong database law. We do not want to break up all our existing markets for intellectual goods just to cure an alleged market failure for investments in a single type of intellectual good, i.e., noncopyrightable collections of information. At present, the U.S. dominates this market,[191] and there is no credible empirical evidence of market failure that could not be cured by more traditional means.

What all this demonstrates is that an exclusive property right is the wrong kind of solution for the database protection problem. Traditionally, information as such was only protected by liability rules – that is, as secret know-how – and not by an exclusive property right. The real need is to devise modern liability rules to protect data that can

---

[189] *See* Heller & Eisenberg. Even without a *sui generis* database regime, privately generated databases distributed online and subject to technologically protected adhesion contracts could bring about the same results, especially if UCITA were uniformly to validate one-sided mass market contracts nation-wide.

[190] Instead of patenting a biotech invention, the investor may seek to maintain a perpetual data base of biotech data that may be difficult or impossible to regenerate. *See* EIPR (2000); Reichman & Uhlir. Over time, the comparative advantage from owning a large database will tend progressively to elevate these barriers to entry. Benkler (2000).

[191] Communication from Maurer (2001), and NRC (1999).

avoid market failure without impoverishing the public domain.[192] The foregoing analysis also demonstrates that, whatever database regime is ultimately enacted, the problems of adhesion contracts and self-help measures will not disappear and must instead be resolved at the same time.[193]

Supporters of strong database protection laws and of strong contractual regimes, such as the Uniform Computerized Information Transactions Act (UCITA), seem to believe that the benefits of private property rights are without limit, and that more is always better.[194] They expect a brave new world in which huge resources will be attracted into the production of databases in response to these powerful legal incentives.[195]

In contrast, critics fear that an exclusive property right in compilations of data, coupled with the proprietors' unlimited power to impose electronic adhesion contracts, will compromise the operations of existing systems of innovation, which depend on the free flow of upstream data and information. They predict a steep rise in the costs of

---

[192] *See* Reichman & Samuelson (1997). This suggests two possible models: old fashioned unfair competition law (sounding in misappropriation), which protects against market destructive conduct or a new form of relief, which may be referred to as a "compensatory liability regime". The latter regime would freely allow second comers to extract protected data for follow-on applications, so long as reasonable royalties were paid to first comers for a reasonable period of time. This approach would solve both the economic and constitutional problems, and would provide the only sound solution to the crucial problem of follow on applications. *See esp*. J.H. Reichman, Of Green Tulips and Legal Kudzu: Repackaging Rights in Subpatentable Innovation. (2000)).

[193] The following measures to regulate licensing seem necessary no matter which database regime is adopted in the end:

1) Sole-source providers must license on fair and reasonable terms.
2) No contractual license can overturn codified exemptions.
3) Courts must have access to some codified criteria of misuse. (*See esp.* H.R. 1858 for specific guidelines in this regard).
4) A general doctrine of "misuse of contracts" or "public-interest unconscionability" should be made available to regulate non-negotiable terms, in mass-market contracts restricting the availability of data as such.

See Reichman & Franklin (1999).

[194] *See e.g*., Raymond Nimmer, cites

[195] EC Directive (Recitals); quoted in Maurer (2001).

information across the global information economy and a serious, long-term "anti-commons" effect that will tend to suffocate innovation by making it contingent on complex chains of contractual permissions that will become necessary simply to procure the basic inputs of the information economy.[196]

In place of the explosive production of new databases that supporters envision, critics fear a progressive balkanization or feudalization of the information economy, in which fewer knowledge goods will be produced as more tithes have to be paid to more and more information conglomerates along the way.[197] In the critics' view, the information economy most likely to emerge from an exclusive property right in data and other pending measures will resemble models already familiar from the middle ages, in which goods flowing down the Rhine River or goods moving from Milan to Genoa were subject to dozens, if not hundreds, of gatekeepers demanding tribute.

2. The Challenge to Science

The point is that the governmental and nonprofit sectors of the modern economy that have heretofore played such a critical role in many national systems of innovation face new and serious threats under these conditions. On the one hand, the research community can join the enclosure movement and profit from it. Thus, universities that now transfer publicly funded technology to the private sector can also profit from the licensing of databases. On the other hand, the ability of researchers to access and aggregate the information they need to produce upstream discoveries and innovations may be compromised both by the shrinking dimensions of the public domain and by the demise of the sharing ethos in the nonprofit community, as these same universities and laboratories see each other as competitors rather than partners in a common venture.[198]

---

[196] *Accord*: Maurer (2001).

[197] All non-profit activities will be especially hard hit. Over time, we predict that lost opportunity costs in neglected research and development projects owing to these balkanized inputs will become staggering, and that many forms of innovation may stagnate as a result. Even so, we will not easily be able to document these lost opportunity costs, and the past experience of science in this regard will be repeated across the whole information economy. *See, e.g., supra* notes _____ and accompanying text (Landsat fiasco).

[198] *Cf.* Eisenberg.

Any long-term solution must accordingly look to the problems of the research communities and of nonprofit users of data generally, in an increasingly commodified information environment. The ability of these communities to oppose or derail what Boyle and Benkler have called the Second Enclosure Movement[199] is limited at best, even if their members were united in such opposition. In practice, the nonprofit communities are divided in their own responses to this movement as they weigh the reduction of government subsidies and their own capacities to commercialize all forms of information technologies, including databases.

The role of the universities and other nonprofit research institutions is critical from this perspective. Universities receive grants of public funds to promote research; they use their own endowments and other funds to conduct research; and they accept privately financed research projects. How universities structure the rules of ownership governing their data and how they regulate inter-university access to their databases will largely determine the availability of data to the scientific community as a whole over time.

If the universities' legal and technology licensing offices formulate these rules and policies from the bottom up, their object will be to maximize the returns on each project without regard to the sharing ethos or to the scientists' need to use data in common. Contracts developed at universities might then resemble those of the private sector, and the profit-maximizing goals of the legal offices would drive the rules applicable to other researchers.[200]

However, experience shows that these narrow revenue enhancing goals soon tend to cancel each other out and lead the universities to impose such mutually unacceptable restrictions on each others' future applications as to bargain to impasse.[201] The combined transaction costs and anticommons effects of each university's contractual regime could gradually make it harder for them to acquire the large amounts of data, from multiple sources, that are increasingly needed for effective research.[202]

---

[199] James Boyle (this conference); Benkler.

[200] For evidence of this trend, see Heller & Eisenberg, *Science*.

[201] *See supra* note ____.

[202] Quote Lederberg. While some universities might experiment with a form of "patent pooling," *see* Merges, this could be hard to organize and would mainly benefit those larger institutions with big packets of rights to trade. It would not help single investigators or small scientific communities

Carried to an extreme, this war of research entities against one another conducted by their respective legal offices could obstruct and then destroy the scientific data commons. As commodification proceeds and intellectual property rights multiply, the functions of the public domain that are now taken for granted may have to be reconstructed contractually by the nonprofit actors engaged on specific projects. Such endeavors could easily fail if different groups seek to overcome rising transaction costs in different ways. If these obstacles to collective action are allowed to grow, moreover, one can foresee endless lost opportunity costs as the scientific community moves away from a sharing ethos.[203]

In previous articles, we have outlined the cumulative negative effects that such tendencies would have on scientific endeavor. For the sake of brevity, we recall them here in summary form:

- monopoly pricing of data and anti-competitive practices by entities that acquire market power, or by first entrants into niche markets;
- increased transaction costs driven by the need to enforce the new legal restrictions on data obtained from different sources, by the implementation of new administrative guidelines concerning institutional acquisitions and uses of databases, and associated legal fees;
- less data-intensive research and lost opportunity costs;
- less effective domestic and international scientific collaboration, with serious impediments to the use, reuse, and transformation of factual data that are the building blocks of modern research.

To avoid these outcomes, science needs to take its own data management problems in hand. The idea is to recreate, by voluntary means, a public space in which the traditional sharing ethos can be preserved and insulated from the commodifying trends identified above. What unites, or should unite all these communities, is a common understanding of the historical function of the public domain and a common need to preserve that function despite the drive for commodification. Although

---

who would face daunting barriers that both the private and public sectors would be creating. In other words, a different and conceptually comprehensive form of pooling is needed. *See infra* Part III.

[203] Documented in NRC, BITS OF POWER (1997); *see also*, Reichman & Uhlir (1999).

legislators and entrepreneurs may take time to understand the threat that a shrinking public domain poses for the national system of innovation, the one group that is best positioned to appreciate that threat is the nonprofit research sector whose dependence on the public domain remains a matter of everyday practice and vital concern. This sector is also the best positioned to take steps to respond to the threat by appropriate voluntary collective action.

It therefore seems advisable for the research community to address these challenges frontally by seeking, of its own initiative, to recreate by consensus and agreement, a dynamic public domain that could ensure a continuous flow of raw materials through the national innovation system, notwithstanding the pressures for commodification from the private sector. In other words, universities and laboratories that depend on sharing access to data will have to stipulate their own treaties and arrangements to ensure unimpeded access to commonly needed raw materials in a public or quasi-public space, even though each institution separately engages in transfers of information to the private sector for economic gain.

This strategy requires a set of rules, standard-form licenses, and organizational structures to be imposed from the top down – by government funders, university administrations, and the leaders of research communities – to institute and maintain a working, dynamic commons in which legal rights are used to strengthen the sharing norms of science along a horizontal public-interest research dimension. At the same time, the rules and norms applicable to this horizontal research dimension must be kept from disrupting the capacities of single actors – universities or researchers – to privatize and exploit their data in a vertical, commercial dimension so that it does not impede public-interest science.

The idea is not to constrain the private domain; it is, rather, to prevent the privatizing ethos from undermining the economically more efficient distribution mechanisms of the sharing ethos in the nonprofit sphere of activities, on which both the public and private sectors depend. If the strategy succeeds, the end result should be to enrich the vertical, commercial domain with more and more downstream applications that emerge from the successful operations of the reconstituted commons. But if nothing is done to preserve the sharing ethos, the risk is that, without a workable functional equivalent of its functions, the horizontal or public-interest sectors would

wither under the pressure of unrestrained, bottom-up commodification efforts. The end result would then be less – not more – innovation.

Fortunately, models already exist that support this general idea of a contractually constructed domain in which intellectual property rights, contracts, and technological measures are used to reinforce norms of sharing for the greater interest of a collaborative community. The open-source software movement, for example, provides basic modalities that support this approach and that have been tested in practice.[204] In this same vein, the idea of constructing a kind of nature conservancy or voluntary "Electronic Commons," or "e-commons," for public access to different types of subject matter is currently under investigation in the United States.[205] In the rest of this paper, we explore the basic concepts that such an e-commons would entail, if applied to scientific and technical data, with a view to convening further workshops to consider this idea. If these initiatives prove successful at the national level, similar efforts would have to be undertaken at the international level as well, in order to extend the benefits of a dynamic e-commons to scientists and other research communities around the world.

III. A CONTRACTUALLY RECONSTRUCTED PUBLIC DOMAIN FOR SCIENCE AND INNOVATION

Some six years ago, when the National Academies first started studying the database protection problem, the authors of this paper first recognized that the mounting pressures on the public domain in scientific data would eventually require science to develop its own new modalities for managing its data supplies.[206] There was, however, no formal model available for easily implementing that objective, and the

---

[204] Berkman Conference papers and materials.

[205] Boyle (DUKE L. J.); discussions with Abelson, Boyle, Lessig, Saltzman (DUKE UNIVERSITY, _____); Benkler workshop (NYU); Berkman Center meeting.

[206] *See* Reichman & Uhlir (1999). That impression was strengthened during the Senate Judiciary Committee negotiations on an early version of the database protection bills in 1998, when negotiators for the scientific and library communities insisted that government-generated data sets, when benefitting from private sector value-adding uses, nonetheless needed to be kept available for public accessibility and especially, for research purposes.

scientific community was not yet sufficiently aware of the deeper problems that an impending assault on the public domain was likely to cause.

A. An E-Commons for Science

Recently, however, considerable thought has been given to the construction of voluntary social structures to support the production of large, complex information projects.[207] Successful implementation of cooperative production and management techniques in regard to the GNU/Linnux Operating System provides one important new model for addressing this problem. The open-source approach adopted by the software research and related communities[208] relies on existing legal regulatory regimes to create a social space devoted to producing freely available and modifiable code.[209]

1. The Basic Concept

Under the GNU/Linnux operating system, components of the cooperatively elaborated structure are protected by intellectual property rights, in this case copyrights, and by licensing agreements, but these legal institutions are used to enforce the sharing norms of the open-source community. Standard-form licensing agreements are formulated "to use contractual terms and property rights to create social conditions in which software is produced on a model of openness rather than exclusion."[210] Under these licenses, "code may be freely copied, modified, and distributed, but only if the modifications (derivative works) are distributed under these terms as well."[211] Property rights are "held in reserve to discipline possible violations of community norms."[212] The end result, as Professor McGowan recently observed is not a true commons, but it resembles a commons because of the "low cost of copying and using code combined with ... broad grants of the relevant licenses."[213]

---

[207] *See e.g.*, David McGowan, Legal Implications of Open-Source Software, 2001 U.ILL.L.REV. 241, 245; *see also* Benkler; Boyle; Froomkin; Radin; Lessig; Berkman Center Papers.
[208] *See* Stallworth and Free-software literature cited in McGowan.
[209] McGowan, 244.
[210] McGowan, 243.
[211] *Id.*, 242.
[212] *Id*., 244.
[213] *Id*., 244.

Recent proposals to launch a "nature conservancy" for information or a voluntary public domain in the form of an Electronic Commons[214] to respond to the mounting threat of an enclosure movement attempt to generalize lessons drawn from the open-source movement and to supply a conceptual framework for thinking about ways to address this challenge. The e-commons concept seeks to reinforce cooperative models for the production of basic information infrastructures in the new knowledge-based economy. The operating principle is that authors, inventors, and other creators can be persuaded to make their works available to the public under General Public Licenses that preserve many of the functions of a public domain without necessarily impeding reasonable commercial uses.[215]

Whatever the merits of this proposal in other spheres of activity, it seems uniquely well suited to the dissemination function of data within the scientific community. We have particularly in mind the need to administer and provide access to databases for scientific research, although, if successful, the e-commons concept could be extended to a multitude of other scientific activities.[216] The real challenge is not just a negative one, i.e., to resist overt, protectionist legislative pressures, such as the proposed *sui generis* exclusive right in databases, or to fashion defensive legal measures against electronic adhesion contracts. Rather, it is to convert the scientific community from errant suppliers and passive consumers of a shrinking public domain to active participants in the construction of a dynamic e-commons, in which the suppliers of public-interest data become technologically linked and accessible in a virtual universal data archive operating for and on behalf of the public interest.

In effect, the scientific community, through its governmental and academic institutions, can reinvent the concepts and function of the public domain in the new technological context. The idea is to construct a new commons space in which the scientific community actively and rationally manages and distributes its own data along a horizontal, not-for-profit dimension. On this horizontal plane, we envision the development of a Linnux-like open system or virtual universal archive in which the participating databases can be accessed and shared for scientific and educational

---

[214] James Boyle, DUKE L. J.; Berkman Papers. *See also* proposals by Bencher and discussion thereof NYU Papers.

[215] Berkman Center Papers.

[216] The publishing of scientific papers and journals is a prime candidate. *See* cites.

purposes under a menu of terms and conditions that the relevant communities themselves negotiate and set in place. The existence of such a horizontal commons, linked by inter-institutional treaties, would preserve access to upstream data for public-interest uses, without unduly disrupting the ability of some community members to commercially exploit their data in a vertical dimension in which commercial applications by the private sector predominate.

2. Differentiating Centralized from Decentralized Suppliers

We do not mean to imply a need to totally reinvent or reorganize the existing universe in which scientific data are disseminated and exchanged. The opposite is true. As we have explained, a vast public domain for the diffusion of scientific data, especially government-generated data, exists and continues to operate, and much government-funded data emerging from the academic communities also continues to be disseminated through these well-established mechanisms.[217]

Centralized facilities for the collection and distribution of government-generated data are well-organized. They are governed by long-established protocols that maintain the function of a public domain and ensure open access and unrestricted use of the relevant data collections. These collections are housed in brick-and-mortar data repositories, many of which are operated directly by the government, such as the NASA National Space Science Data Center or the National Center for Biotechnology Information at the NIH. Other repositories are funded by the government to carry out similar functions, such as the archives of the National Center for Atmospheric Research (NCAR) or the Hubble Space Telescope Science Institute.

Under existing protocols, most government-operated or government-funded data repositories do not allow the conditional deposits that look to commercial exploitation of the data in question. Anyone who uses the data deposited in these holdings can commercially exploit their own versions and applications of them without needing any authorization from the government. However, no such uses, including costly value-adding uses, can remove the original data from the public repositories. In this sense, the value-adding investor obtains no exclusive rights in the original data, but is allowed to protect the creativity and investment in the derived information products.

[217] *See supra* text accompanying notes _____.

The ability of these government institutions to make their data holdings broadly available to all potential users, both scientific and other, has been greatly increased by direct online delivery and telecommunications networks. However, this potential is undermined by a perennial and growing shortage of government funds for such activities; by technical and administrative difficulties that impede long-term preservation of the exponentially increasing amounts of data to be deposited; and by pressures to commodify data, which are reducing the scope of government activity and tend to discourage academic investigators from making unconditional deposits of even government-funded data to these repositories.[218]

The long-term health of the scientific enterprise depends on the continued operations of these public data repositories and on the reversal of the negative trends identified earlier in this paper. Here the object is to preserve and enhance the functions that the scientific commons has always played, notwithstanding the mounting pressures to commodify even government-generated data.

At the opposite extreme, ever-increasing amounts of important scientific data, including both government-generated and government-funded data are controlled by an anarchical structure of highly distributed individual investigators or small teams of investigators. Operators at this end of the spectrum do not rely on large research facilities to conduct their investigations and they generate their own relatively small and heterogenous data sets, which they maintain autonomously. Examples may be found in biotechnology, biomedical and biodiversity research, ecology, and the behavioral sciences, among many others.

Under this decentralized model, the protocols for depositing data in public-domain repositories or for making them otherwise available are typically less well developed or nonexistent. Because the data are controlled by autonomous investigators, they have considerable freedom of action, they encounter few mandatory requirements, and there are ways to avoid any disclosure requirements that funders may impose. Furthermore, much of the data in this category tends to be contractually restricted and can become proprietary, and in certain areas of great commercial promise, there are strong pressures not to make the data available on a limited basis even to other academic or nonprofit investigators. In other areas, such as biomedical

[218] *See supra* text accompanying note _____.

and behavioral research, there are additional restrictions grounded in privacy and confidentiality concerns that need to be respected in any reform efforts that may be undertaken to ensure greater access to data generally.[219]

Apart from these pressures and problems, the ability of single investigators to participate in a collaborative structure and to share their data broadly was further limited until recently by traditional modes of dissemination in print media. The advent of digital technologies and the Internet, however, have made it possible to integrate even the data outputs of single investigators and their small communities on a cooperative basis. In other words, the technical means exist to convert previously decentralized and autonomous data collection activities into virtual data centers or "collaboratories" that could mimic the functions and provide many of the benefits of the bricks-and-mortar data centers.

To be sure, one must not assume that these autonomous investigators are all imbued with the sharing ethos that underlies the culture that surrounds and is institutionalized in the large, facility-based research facilities. Indeed, some of these subcommunities have tended to hold onto their data by tradition, even in the absence of economic pressures to commodify, because the sharing ethos was largely extraneous to their fields of endeavor. Ecological and anthropological field studies provide examples, and some economic research also fits here. An added cultural factor in some of these fields is that the academic journals do not require disclosure or public deposits of underlying data at the time of publication.

The point is that this adverse culture needs to be changed in order to take greater advantage of new technological opportunities and their positive network externalities, as well as to broadly disseminate data to resist growing pressures to restrict access in the interest of commodification. Here efforts should be made to promote and expand the sharing ethos embodied in the principle of "full and open exchange of data" to the decentralized players, and to encourage and reinforce their use of open, digitized networks by appropriate legal mechanisms that implement that ethos. This is especially true given that most of the distributed players are academics funded by government in at least some stage of their research.

---

[219] *See supra* text accompanying notes _____.

Unless these investigators are integrated into a larger cooperative system based on the sharing ethos, we shall face the anomalous situation in which government-funded data escape all the distributive functions of a public domain merely because most of the work in question is performed outside government and subject to growing commercial pressures, including such government-initiated measures as the Bayh-Dole Act. If present trends outlined in Part II continue unabated, ever-increasing amounts of scientific data, including publicly funded data, will be removed from public-domain distribution mechanisms and placed within private distribution mechanisms that condition access on the payment of money and that otherwise greatly restrict the secondary uses that can be made of even data that are lawfully accessed.

B. Implementing the E-Commons Approach

To facilitate the exposition of our proposals, we subdivide the concept of a reconstituted public domain for scientific data into two broad categories. In the first category, which is a "pure" public-domain environment, data are deposited or made available unconditionally, and they cannot be removed or become subject to exclusive private ownership. Almost all of the data circulating here will either be government-generated or government-funded.

In the second category, which is conceived as an "impure" or hybrid environment, data are deposited conditionally, and private, exclusive uses are permitted. We envision these private exchanges as occurring along a vertical axis descending from the depository entity, largely (though not wholly) unregulated by the commons regime. However, the public-use licensing and other mechanisms that make the e-commons operational would preserve access to, and encourage sharing of the data deposited in the "impure domain" for research purposes on a trans-institutional basis. We see the mostly nonprofit research entities or investigators who exploit the favorable terms and conditions imposed by the standard-form licenses governing these conditional deposits as constituting a "horizontal" research axis, whose operations are contractually insulated from the legal regimes that govern private-sector transactions occurring on the vertical axis.

1. Instituting an Unconditional Public Domain

Where no significant proprietary interests come into play, the optimal solution for government-generated data and for data produced by government-funded research is a formally structured, archival data center also supported by government. As we have seen, many such data centers have already been formed around large-facility research projects. The first idea we put forward here is to extend this time-tested model to highly distributed research operations conducted by single investigators or teams of investigators. An established example of a data center along these lines is the National Center for Biotechnology Information. Reportedly, the ecology community is also

considering such a center to meet their research data needs. We believe other discipline-specific communities could benefit from similar arrangements.

This proposal is, of course, a prescription for extending the pure public domain concept from its brick-and-mortar origins organized around large central research facilities to the outlying districts and suburbs of the scientific enterprise. It is meant to reconcile practice with theory in the sense that most of these investigators are academics funded by government anyway. By overcoming inertia and ensuring that the resulting data are effectively made available to the scientific community as a whole, the social benefits of public funding are more perfectly captured and the sharing ethos is more fully implemented.

Because unconditional deposits occur in a pure public domain environment removed from proprietary concerns, and there is no vertical axis of commercial or proprietary interests to take into account, the legal mechanisms to implement these expanded data centers need not be complicated. Single researchers or small research teams could contribute their data to centers serving their specific disciplines with no strings attached. Alternatively, as newly integrated scientific communities organize themselves, they could seek government help in establishing new data centers that would accept unrestricted deposits on their behalf.[220]

We note in this connection that many academics have themselves self-organized mini "data centers" through their Web sites with public-domain functions, limited only by their technical and financial capabilities. Groups of academics can similarly construct more ambitious mini-centers, which could become less elaborate versions of the government data center model.

Private companies can also contribute to a pure public domain model, or they can organize their own variants of such a model, and these practices should be encouraged as a matter of public policy. For example, private companies have contributed geophysical data sets from proprietary oil exploration research to government data repositories open to the public. Similarly, proprietary Landsat data have been provided to the U.S. Geological survey's EROS Data Center archive and placed in the public domain after ten years.

---

[220] [cites]

If the unrestricted data are deposited in federal government sponsored repositories, existing federal information laws and associated protocols will define the public access rights. If, however, data centers are formed outside the scope of direct government control, the organizers and managers will need to reconstruct the public domain through general public use licenses to emulate the protocols that govern deposits of data in more traditional government-operated centers. A primary concern here (as in the second or "impure" category discussed below) is to ensure that academics receive suitable attribution and recognition for their data-related activities. There is evidence that one reason open-source software systems have succeeded is that they confer reputational benefits on their participants.[221]

A major stumbling block for creators or operators of data centers (broadly defined) that open-source software communities seem able to avoid is the need for considerable funds to maintain databases over time and to manage the data holdings so as to fulfill the public access functions. This is why government support appears to be indispensable, as an integral part of promoting basic research as a public good. The maintenance of public-interest data centers is problematic without such support. Conceivably, some of these data centers could become partly or fully self-supporting through some appropriate fee structure,[222] but the temptation to restrict subsequent uses must be resisted under such a paying public domain concept. In any event, resort to a fee structure based on payments of more than the marginal cost of delivery quickly begins to defeat the public good and positive externality attributes of the system, even absent further use restrictions.

Assuming the financial hurdles can be overcome, the new digital and telecommunication technologies, coupled with new legal models, create exciting possibilities for constructing totally decentralized or virtual data centers that could facilitate peer-to-peer exchanges of single data products or sets. These exchanges would require appropriate General Public Licenses, and there would need to be at least some minimal administrative structure.[223]

---

[221] McGowan, ____ .

[222] The National Oceanic and Atmospheric Administration's (NOAA) National Data Centers operate along these lines.

[223] This would depend on a number of factors, and may not be required at all.

Scientists, of course, already accomplish such exchanges informally among themselves under the norm of "full and open exchange," but the idea here is to formalize that process and give it a sound legal and organizational framework. This framework is needed for both negative and positive reasons. It would initially help scientists to resist proprietary pressures, including those emanating from the universities,[224] and encourage the placement of data in a true commons, while the existence of GPLs supported by the scientific and funding communities would reduce the legal uncertainties that may inhibit sharing. In a larger perspective, the goal is to facilitate cooperative access and use of multiple sources of data in a more efficient institutional framework that exploits the network externalities the Internet makes possible and that enables the scientific community (and the innovation process) to devise maximum value from the taxpayers' investment in these public-good resources.[225]

This proposal is facilitated by the initial assumption that the relevant data will be deposited unconditionally and without encumbrances or restrictions, other than perhaps certain requirements concerning attribution (a form of moral rights). Needless to say, this excludes a large and growing sector of scientific endeavor whose data outputs cannot, for various reasons, be unconditionally deposited in a true commons. For this sector, we must contractually construct a less pure version of a commons that would reconcile the competing interests of open access and use for research purposes with commercial exploitation.

2. Conditional Public-Domain Mechanisms

Candor requires us to admit at the outset that U.S. science policy disfavors a two-tiered system of data distribution.[226] While we sympathize with the philosophy behind this position, our six years of focused study on issues concerning the legal protection of databases[227] compels us to consider the realities of a growing trend toward two-tiered distributive activities in order to determine whether such activities

---

[224] Rai & Eisenberg (2001).

[225] Citing Stiglitz, *supra* ___.

[226] Such two-tiered systems for government or academic data distribution have been favored and promoted by the scientific community in the E.U., and these initiatives have been strongly opposed by U.S. science agencies and academics.

[227] *See* NRC 1997; NRC 1999; Reichman & Uhlir (1999); Reichman & Samuelson (1997); *see also* Uhlir (1998), (1999), (2000), and Reichman (publication pending).

can be operated in a manner that preserves the benefits of a public domain, notwithstanding the mounting pressures for commodification.

European governments have already embarked on a policy of commercial exploitation of publicly generated data and even insist on conditional deposits in various governmental scientific organizations and in cooperative research activities. Some academic scientific communities have recently tried to commercialize biotechnology databases of considerable public research value on a two-tiered basis,[228] while others have succeeded with controversial results.[229] The reality is that U.S. universities intend to commercialize some of their data and support minimalist legislation to this end. Conversely, some enlightened and promising private firms, such as Celera Genomics and [Minnesota GMO firm],[230] have made their expensive databases conditionally available to the scientific community on favorable terms, and such initiatives to maintain access to a scientific commons for nonprofit researchers should be encouraged.

a. Characteristics of an impure domain

In this domain, owners of databases envision split uses of the data and will only make them available on restricted conditions. Some of these uses are for pure research purposes in nonprofit entities, while others entail purely commercial applications. Moreover, these two zones of activity are not neatly or clearly separable, which adds to the costs and complications of administration. For example, universities may treat some databases as commercial research tools with a price discrimination policy that provides access to the research community at a lower cost than to for-profit entities.

In the impure domain, the funding of data production is generally (but not always) less dominated by government, with more of the financial burden borne by the research entities themselves, especially by universities, by private companies, or by cooperative research arrangements between universities and the private sector. Despite their educational missions and nonprofit status, the universities are increasingly prone to regard their databases as targets of opportunity for commercialization.

---

[228] Hugo Mutations Database. See Maurer (2001).
[229] See Swiss PROT; *see also* NRC Study (1999).
[230] See address by CEO at Minnesota GMO Conference.

Scientific data can also be made conditionally available in an "impure domain" through complicated three-way funding arrangements typically initiated by government science agencies under CRADAs. Complications in this instance arise from tensions between the government's continued interest in promoting public access and legislative policies, as embodied in the Bayh-Dole Act, which encourage commodification of government-funded research results. Even here, however, the fact that the government's financial contribution to the project may predominate gives it the clout to impose conditions favorable to public-interest research uses. At present, this power is under-utilized,[231] but a major purpose of establishing a solid legal framework for conditional deposits would be to provide standard-form licenses that clearly reinforce and implement favorable public-interest terms and conditions, without unduly compromising the commercial interests.

With these factors in mind, our second major proposal is to establish an impure zone of conditionally available data in order to reconstruct and artificially preserve functional equivalents of a public domain. This strategy entails using property rights and contracts to reinforce the sharing norms of science along a nonprofit, trans-institutional (horizontal) plane, without unduly disrupting the commercial interests of those entities that choose to operate in the private (vertical) plane. This project presupposes a formal understanding among the major players analogous to "multilateral treaties," particularly, the government's science funding agencies, the universities, and the scientific community, and it would benefit greatly from collaborative arrangements with for-profit research entities in the private sector.[232]

We recognize that an impure domain of conditionally deposited data is, for many purposes, clearly a second-best solution.[233] However, unless such a zone is set in place with the express goal of preserving access to data for public-interest uses, the pressures for privatization and commercialization may be carried so far as to subject most public uses to "private ordering" under intellectual property rights, adhesion contracts, and technological fences.[234] One should thus conceive of the impure domain

---

[231] *Cf.* Rai & Eisenbergi, in the context of biotech patents.

[232] Possible antitrust implications would need to be addressed.

[233] *See supra* for the optimum solution, i.e., unconditional deposits in public-domain data centers.

[234] Lest we be deemed hyperbolic Cassandras, we reiterate that such a regime has become increasingly common in other countries.

as a buffer zone that preserves and expands the social benefits of a commons, despite the pressures to commodify scientific data.

We also recognize that an impure zone poses administrative complications, costs, and other drawbacks. Clearly, the allowance of restrictions on use breaks up the continuity of data flows across the public sector and necessitates burdensome administrative measures and transaction costs to monitor and enforce differentiated uses. It also entails measures to prevent unacceptable leakage between the horizontal and vertical planes, and it may result in changes that exceed the marginal cost of delivery for public-interest uses on the horizontal plane.

The inescapable conclusion is that the impure domain dilutes the sharing ethos and constitutes an option of last resort. As we read the tea leaves, however, the enclosure movement appears to be advancing inexorably. The only way to preserve and reinforce the sharing ethos of science in a new world of increasingly commodified scientific data is to appropriately implement this option of last resort. With these premises in mind, we envision three specific situations in which the desirability of a two-tiered approach needs to be considered: 1) the public sector, 2) the academic environment, and 3) the private sector. In the following sections, we will suggest that a two-tiered approach is, in fact, undesirable for public-sector activities; that it has become a necessary feature in the academic environment; and that it is highly desirable in private sector undertakings.[235]

b. Sectoral Evaluations

(i) The public sector

Everything we have written in support of the pure domain of unconditional deposits and availability shows why a two-tiered approach is highly undesirable with respect to government-generated data. The American tradition is squarely opposed to restricted uses of such data. However, many European and other governments (including the U.K. and Canada) have subscribed to a different tradition, and the European Union's Database Directive represents a powerful new thrust in that

[235]We limit our proposals here to a general conceptual framework. A subsequent article will elaborate them in more detail and provide examples.

direction. This model enables governments to exercise strong and perpetual exclusive rights in publicly generated databases, without any mandated obligation to recognize public-interest exceptions.[236]

Some fifty states that either belong to the E.U. or have an affiliated status are expected to adopt this model, and E.U. trade negotiators have sought to impose it on other countries as part of regional trade agreements. If the United States fails to adopt a different, less protectionist database regime, founded on true unfair competition principles, the pressures for other countries to follow the E.U. Directive will be very great. Even if the U.S. adopts a significantly less protectionist model, however, there will be pressures on the U.S. to protect data generated by foreign governments that are made available to data centers in the U.S., despite the no conditional deposit rules that bind many of these centers. The U.S., of course, will not be able to prevent foreign governments from commercially exploiting their public data in territories governed by the E.U. Directive. On the contrary, the fact that governments in the E.U. themselves saw this Directive as a source of considerable income most likely disposed them favorably toward it, and this fatal attraction seems to be spreading.[237]

For these reasons, and despite the general undesirability of a two-tiered structure in the public sector, it is indispensable that governments that choose to exercise crown rights (both copyrights and *sui generis* rights) under the E.U. Directive or its analogues in other countries take steps to implement an "impure" domain,[238] with a view to maximizing access for nonprofit research, education and other public-interest purposes. At the same time, there is a real danger that the E.U. will press many intergovernmental organizations, as they have the World Meteorological Organization (WMO) and the International Oceanographic Commission (IOC) already, to adopt two-tiered systems that deviate from established U.S. norms and policies. The E.U. has also pressed U.S. government agencies to conditionally protect the former's data in intergovernmental exchanges and thus, in effect, to institute a two-tiered approach for some purposes at U.S. data centers. Similarly, the E.U. has pressed the U.S.

---

[236] *See supra* text accompanying notes _____.

[237] See, e.g., the case of Korea. For the moment, Canada, Japan, and other OECD countries are sitting on the fence. *See* Maurer; [Japan cites].

[238] Obviously, the better result would be for the E.U. governments to renounce crown rights and to adopt the full and open policy of the U.S. government. Some efforts in this direction are underway, but the outcome is highly uncertain (cites).

government to retreat from its "full and open" data exchange policy in international scientific research programs, and it appears they have sometimes succeeded in obtaining restrictions on access to, and use of, data beyond the immediate research objectives.

As we stated at the outset, a two-tiered system is antithetical to the information policies that traditionally regulate government-generated data in the U.S. It also conflicts with established U.S. science policy and with the economics of the public domain. There are thus many reasons for characterizing the European approach as both backwards-looking and counterproductive, with negative implications for both scientific cooperation and local innovation. Nevertheless, if nothing persuades the E.U. to change its present direction, or if similar pressures are successfully applied to the U.S. government, new adjustments may be needed to the existing "pure" domain for the distribution of government data, at least at the international level and possibly even in the U.S.[239] In that most regrettable case, the only way to preserve and enhance the social space for U.S. government-generated data may be to adopt the two-tiered variant discussed below. Naturally, we continue to hope this option will not become necessary and we do not further explore it in this paper.

(ii) Academic sector

Scientific database production in academia is not necessarily dominated by government-funding and may entail funding by universities, foundations, and the private sector. Nevertheless, it is well to remember that public funding remains a presence in this sector, and its role varies from project to project.

The solution we envision here is to maintain the functions of a public domain to the fullest extent possible on a horizontal level that provides access for nonprofit research activities, and to encourage efficient technological uses of the data available in

[239] As noted in Part II, some industry groups are pressing the U.S. government to transfer the data dissemination function to the private sector in order to capture it. Others want the U.S. science agencies to stop generating their own data and to license the data from the private sector. Meanwhile, the E.U. presses the U.S. to adopt their two-tiered structure, and the denial of national treatment in the Directive reinforces these pressures.

this domain. At the same time, commercial exploitation under more restricted conditions would be permitted on the vertical plane.

Linking the Communities

This solution provides both negative and positive benefits. Negatively, the object is to preserve a public space and the efficiencies it makes possible from encroachment by the "do it our own way," profit-maximizing mentality of university technology licensing offices and of other commercializing initiatives. Unless steps are taken to parry the tendency of each actor to impose its own terms, without regard to the interests of the research community, there is a risk that data-intensive research activities will lapse into balkanized private zones, in which exchange and innovation are impeded. A well-documented example of how this can occur is to be found in the Human Mutations Database Initiative, where failed efforts to commercialize a collection of independently generated, highly valuable databases while preserving public-domain access for the nonprofit researchers themselves, has left the collection in precisely this kind of balkanized state.[240]

On the positive side, our proposed solution presents an opportunity to institute and enlarge new public-domain-like zones whose functionality can be potentiated by digital network technologies. As discussed in Part I, academic researchers or research teams in the past have not necessarily made their data available to others, particularly in highly distributed, "small science" research areas. Even where a desire to do so may have existed, there were technical limitations and legal uncertainties in the way, as well as a risk of depreciating the commercial value of the data in question. Moreover, funds to promote sharing, or some institutional structure to support it, may well have been lacking, and the practice of sharing brought no certain reputational benefits.

The e-commons concept turns these difficulties around and makes a virtue out of necessity. It allows both single researchers and small communities to link up technically, to share access to data at the very least, and possibly to co-administer their data holdings in the common interest. Indeed, these improved linkages could themselves become a bourgeoning and productive source of data that might otherwise have been left untapped for lack of appropriate mechanisms. Here we have in mind the possibilities for productive gains that can be realized from interdisciplinary and cross-

[240]Maurer (2001).

sectoral uses, and also from cooperative management techniques roughly analogous to some of those used in, say, the open-source software movement.

To construct such a two-tiered e-commons solution, however, many obstacles must be overcome. Initially, the very concept of an e-commons needs to be sold to skeptical elements of the scientific community whose services are indispensable to its development.[241] Academic institutions, science funders, the research community, and other interested parties must negotiate and stipulate the pacts needed to establish an impure domain as well as the legal framework to implement it. Transaction costs will need to be monitored closely and, whenever possible, reduced throughout the various development phases.

Universities will also have to be sold on the benefits of an e-commons for data, with a view to rationalizing and modifying their disparate licensing policies, which often seem as or more restrictive than those of their private-sector counterparts.[242] This project will require statesmanship, especially on the part of the leading research universities, and it may require pressure from the major government funders of the universities to encourage them to develop agreed and appropriately varied General Public Licenses. Account will have to be taken as well of the universities' patenting interests, which will need to be suitably accommodated.

Here recent experience with the open-source software movement provides some useful models, but it also suggests certain constraints that the scientific data construct will have to face. Clearly, the success of the open-source software movement provides a positive model in so far as it indicates the potential gains flowing from standardized licensing agreements and from the use of both property rights and contract to enforce community norms, in this case, the sharing ethos. It is also a good model for producing the reputation rewards,[243] which we deem essential to the success of this initiative.

---

[241] See criteria set out in personal communication from Harlan Onsrud.
[242] Heller and Eisenberg; another Eisenberg; NIH report.
[243] *See* McGowan.

However, the open-source software movement has generally shied away from trading access for payment,[244] and tends instead to shunt all activities involving payments to the private or vertical dimension, in which some firms, notably Red Hat, have flourished.[245] While this practice seems desirable, it may not be transplantable to the realm of scientific databases.

For one thing, universities regard some databases as research tools, which, even if not patented, they will want other universities to pay to use. Moreover, the relatively high cost of preserving and maintaining data holdings under present-day conditions may make a certain financial return from providing access indispensable even along the horizontal axis. In this regard, there is ample reason to believe that public funds would not be adequate to support the costs of managing all needed activities in the pure zone, much less the impure zone as well, even if universities and funding agencies could otherwise agree on the appropriate legal and administrative structure to implement the e-commons concept. In other words, even if the universities' profit-maximizing inclinations are satisfactorily moderated, there most likely remains a built-in need to collect at least part of the costs of managing and archiving the data holdings from participating users. While government support ought to increase, especially as the potential gains from a horizontal e-commons become better understood, the costs of data management will also increase with the success of the system. For this reason, it would be necessary at a minimum to levy charges against users in the private sector who operate in the vertical dimension, in order to help to defray the costs of administering operations in the horizontal domain and to make this overall approach economically feasible.

While pressures to extract payment for reasons other than defraying management costs should be resisted, especially if a preponderance of funding comes from government sources, the need to cover management and related transaction costs is a reality that one cannot ignore. We have recognized that charges levied for use of data in the impure domain would have to take into account the costs of data management, unless otherwise defrayed by government, although we hope that the bulk of these costs could be recovered from private-sector uses on the vertical plane, rather than nonprofit uses on the horizontal plane.

---

[244] Id.

[245] Id.

A realistic appraisal of current practices nonetheless compels us to examine potential demands by academic suppliers for payments in excess of data management costs to be levied even against nonprofit users on the horizontal plane. At present, such demands are likely to occur for users of what are perceived to be so-called "research tools," and they are often accompanied by onerous contractual conditions, especially clauses seeking to establish reach-through claims on follow-on applications obtained by value-adding users, for-profit and nonprofit alike.

Compensatory liability

Whether negotiations leading to a multi-institutional "treaty" establishing an e-commons could altogether eliminate or regulate such demands and legal modalities remains to be seen. Assuming that a peace pact cannot completely remove the underlying concerns that prompt "reach-through" and similar demands, our preferred solution is to mandate a compensatory liability approach to follow-on applications[246] that would at least deny suppliers any hold out or veto rights over value-adding uses by lawful participants on the horizontal plane.

A compensatory liability mechanism could allow certain restricted uses of certain agreed kinds of data for certain agreed purposes (e.g., follow-on applications of specified research tools) by participants in the horizontal, nonprofit dimension in return for reasonable contributions to the costs of developing, maintaining, and servicing the data holdings over a specified period of time. These payments, if allowed at all, should vary with the status of the user. Moreover, an indispensable condition of such a regime is that any academic supplier who provides data for follow-on applications subject to the compensatory contribution mentioned above should also benefit from an absolute right to borrow back the second academic comer's value-adding contributions, for research purposes, subject to similar compensatory liability payments for a similarly reasonable period of time.

In effect, a compensatory liability mechanism eliminates the possibility of academic suppliers imposing vetoes and hold-out options, including reach-through clauses, on other academic entities that develop follow-on applications using the conditionally available data. At the same time, the compensatory liability mechanism

[246] *See*, J. H. Reichman, Green Tulips; Legal Hybrids.

would make all the participants in the "paying public domain" segment of the impure zone into a *de facto* cooperative group for purposes of certain agreed value-adding applications of the common data holdings, and it would not allow any academic participant to impose "exclusive rights" options that extracted payment for use at the cost of impoverishing the contractually reconstructed e-commons.

Administrative considerations

Looking beyond these troublesome, but unavoidable, questions of payment for research uses at the margins of the horizontal dimension, there are questions about how technically to organize the impure zone as a whole that would have to be resolved. For example, the peer-to-peer file-sharing solution that we discussed earlier in this paper,[247] would presumably still provide satisfactory results if used to link different scientific communities participating as such in the pure zone. However, there are reasons to doubt that it can produce equally satisfactory results when linking single investigators operating in the impure domain, each of whom remains the master of his or her own data for all purposes.

The better solution may be for participating investigators in this zone to deposit their data with an administrative agency or service charged with the task of supplying and administering the General Public Licenses, subject to the guidance, governance, and oversight of an appropriate body in which government funders, universities, and other relevant institutions were represented. If such an administrative service could be established on a solid footing, it might then become feasible for it to provide Napster-like linkages among single data suppliers, even under a totally decentralized approach.

Implicit in these considerations is the larger question of how to develop, promote, and enforce the General Public Licenses needed to render the impure zone operational without some hierarchical administration that would perform functions analogous to those that Linus Torvald performs with respect to the GNU/Linnux Operating System.[248] In principle, a private, voluntary group, such as the Berkman Center's e-commons group, could perform these functions, but we anticipate that the scientific community would itself eventually want to take over some, if not all, of these

[247] *See supra* .
[248] *See* McGowan, *supra*, at .

functions. The logical organizational locus for such operations would be the professional scientific societies working within the framework of the American Association for the Advancement of Science. At the same time, we could also foresee – as indicated above – a bifurcated organizational solution in which an external administrative agency performed the daily functions, including the clearing of rights (on a voluntary basis), subject to oversight and governance by an appropriate scientific entity.

How best to enforce the General Public Licenses and the community norms they support is an integral part of the organizational issues raised above. Clearly, once a service provider – an administrative agency – became proficient, its skills would be attractive to participating communities which, over time, might otherwise have to duplicate these transaction costs with less ability. Hence, we think the administrative agency could become a voluntary clearing house for rights management and for the collection of any payments or royalties from either the horizontal or the vertical sectors.[249] We also envision the need for dispute mediation and dispute settlement facilities, which would be appropriately located in any oversight group that might be established.

Returning for a moment to the thorny problems of payments for research uses even along the horizontal dimension, there is a further disciplinary or enforcement problem arising from the need to avoid leakage of data supplied at preferential prices to research users in ways that might damage the interests of private-sector users in the vertical dimension. It will be recalled that, on the horizontal plane, the option to charge for research uses (when otherwise unavoidable) is intended to entail a corresponding burden positively to discriminate in favor of science and its research goals. This practice is, of course, further restrained to the extent that the U.S. government provides the bulk of the data in the pure domain,[250] which intrinsically restricts the amount of data available for providers who seek to opt into the impure domain, with price-discriminated operations along the horizontal research plane in addition to commercial operations at full rates along the vertical axis.[251]

---

[249] The Harry Fox model under copyright law, as applied to recordings of music, is an interesting example. *See* ; Merges.

[250] *See supra* text accompanying notes .

[251] *See* NRC Study.

This need for price discrimination favoring research uses along the horizontal axis requires that the difficult problem of leakage be addressed. Any solution here would probably require the administrators of a scientific e-commons to adopt and apply its own version of digital rights management techniques with a view to implementing and enforcing the community's norms. Congressional enactment of a minimalist database protection right along the lines of H.R. 1848[252] might help facilitate a solution to this problem of leakage.

Finally, care must be taken to reduce friction between the scientific data commons as we envision it and the universities' patenting practices under the Bayh-Dole Act. For example, the GPLs might have to allow for deferred release of data even into the pure domain, at least for the duration of the one-year novelty grace period during which relevant patent applications based on the data could be filed.[253] Other measures to synchronize the operations of the e-commons with the ability of universities to commercialize their holdings under Bayh-Dole would have to be identified and carefully dealt with in the applicable GPLs.

One consequence of the Bayh-Dole Act in conjunction with our proposed approach could be to encourage data providers to avoid unconditional deposits of data into the pure zone, even when they result from government funding, in favor of deposits to the impure zone, which retain the capacity for restricted uses and some forms of commercial exploitation. Here, however, there are no comparable problems of reconciling a horizontal public access dimension along the lines described above with a vertical, Bayh-Dole dimension. On the contrary, it may be that the presence of a government-funded component could make it easier to control and limit the kinds of restrictions on public access for research purposes that would be permissible under the GPLs that regulate activities on the horizontal planes, without disrupting the policies of Bayh-Dole with respect to activities in the private sector.

We also note that there is an interface between our proposals for an e-commons for science and antitrust law that would at least require interaction with the

---

[252] *See supra* notes ___ and accompanying text.
[253] *See* 35 U.S.C. §102.

Federal Trade Commission and might also require enabling legislation. A detailed anlysis of these issues lies beyond the scope of this paper.[254]

(iii) The private sector

Data funded by the private sector are logically subject to any and all of the proprietary rights that may become available, as surveyed earlier in this paper. Here the object of an e-commons approach is to promote voluntary contributions to the impure domain that might not otherwise become available for research purposes on favorable terms and conditions.

The existence of the e-commons, suitably armed with appropriate GPLs (roughly analogous to the "Lesser General Public Licenses"of the open-source software movement),[255] would thus enable enlightened private-sector research organizations to continue to supply data to a contractually constructed public domain, in exchange for their own abilities to access and use the holdings of public access commons for for-profit research activities. The end result would provide both the scientific research communities and the for-profit research communities of enlightened private-sector participants with access to a more comprehensive, cooperatively maintained data universe on the horizontal plane than would otherwise be possible if access to data for research purposes were to be governed by an excessively rigid distinction between nonprofit and for-profit research endeavors.

The relevant licenses would have to be carefully drawn, however, and we frankly concede that the legal solution might entail a "Much Lesser GPL" variant of a kind unknown to, say, the open-source software community. Moreover, whereas the object in most applications of the e-commons concept we have been discussing would be to rely on nonnegotiable standard-form contracts, relations with the private sector might benefit from more tailor-made variants to accommodate specific firms or particular situations. For example, the need to reconcile Celera's interest in retaining rights to its data while still publishing its genomic results in Science gave rise to the kind of accommodation that might be necessary in other cases.[256] Another case in point

[254] [add section illustrating potential of our system to solve problems that H____ could not resolve]
[255] *See* McGowan.
[256] *See also* Minnesota Genome firm interested in preserving public access.

might be the genomic databases that some pharmaceutical companies have established defensively, with a view to limiting the scope for competitors to obtain patents in specific areas of investigation.[257]

The GPLs applicable to private firms operating in the vertical dimension who opt into a public access commons arrangement could be fairly restrictive in their allowable uses, as compared with the conditions applicable under the GPLs implementing any of the other options discussed above. But the goal of securing greater access with fewer restrictions to privately generated data justifies this approach because it makes available to the research community data that would otherwise be subject to commercial terms and conditions in a more research- unfriendly environment.

There is, of course, a risk that universities would sooner or later see themselves as more like option 3 private players, than like the option 2 players with whom we wish to identify them.[258] This would conflict further with the public mission of the universities, not to mention their tax exempt status. Nevertheless, concerted efforts must be made at the "treaty-making" phase to prevent or discourage the universities from taking this route, and individual investigators, especially academic investigators, should themselves press the universities to adhere to "second option" status. While we concede that efforts to restrain the universities in this regard carry no guarantee of success, the risk that some universities may gravitate toward private sector status in some circumstances seems nonetheless preferable to current practices and tendencies, which are characterized by profit-maximizing, technology licensing officers bargaining to impasse in a commercialized environment that takes little or no account of the need for, and functions of, a public domain.

Another set of problems hinges on possible conflicts of interest between universities and their scientists. In a relatively benign form, such a conflict could arise when the scientists and their teams or communities opt for one set of GPLs, while their universities are inclined toward another. As we implied earlier, the GPLs normally applicable to academic investigators would presumably be more mechanical and group-oriented than the licenses available to private-sector participants, and there would be more room for tailor-made conditions under the latter. Private-sector GPLs would also

---

[257] *See* ___.

[258] *See, e.g.*, Heller & Eisenberg, on the tragedy of the anti-commons.

presumably carry higher transaction costs and, in general, impose more restrictive conditions on access to data for research purposes.[259]

A much bigger set of problems arises when a university sees other scientists as a target market for the research tools it produces. In this situation, it has the same potential commercial interests as private producers of tools for scientific research. Nonetheless, and disregarding moral questions relevant to their academic missions, the universities as a group share a common interest in reducing their overall transaction costs, which conflicts with their individual interests in commercially exploiting selected research products. We trust that the common interest in reciprocal access at acceptable rates would provide a basis for a negotiated compromise, including, where necessary, the possibility of compensatory liability provisions in some cases. We hope that such a compromise could be worked out during the "treaty-making" phase that would have to precede the formation of an e-commons for science along the lines proposed in this paper.

[259] There is little available experience with standard licenses of this type, although some help may be derived from Joint Cooperative Research Agreements.