ARE WE LEGISLATING AWAY OUR SCIENTIFIC FUTURE? THE DATABASE DEBATE

Dov Greenbaum[1]
Yale University, Department of Genetics
dov.greenbaum@yale.edu

The ambiguity of the present copyright laws governing the protection of databases creates a situation where database owners, unsure of how IP laws safeguard their information, overprotect their data with oppressive licenses and technological mechanisms (condoned by the DMCA) that impede interoperation. Databases are fundamental to scientific research, yet the lack of interoperability between databases and limited access inhibits this research. The US Congress, spurred by the European Database Directive, and heavily lobbied by the commercial database industry, is presently considering ways to legislate database protections; most of the present suggestions for legislation will be detrimental to scientific progress. The author agrees that new legislation is necessary, but not to provide extra-copyright protections, as database owners would like, but to create an environment wherein data is easily accessible to academic research and interoperability is encouraged; yet simultaneously providing database owners with incentives to produce new databases. One possibility would be to introduce standardized compulsory licensing of databases to academics following an embargo period where databases could be sold at free-market prices (to recoup costs). Databases would be given some sort of intellectual property protection both during and after this embargo in return for a limiting of technical safeguards and conforming to interoperability standards.

I. INTRODUCTION

¶1 The Bayh-Dole Act of 1980[2] attempted to ease the tense relationship between academia and industry over information sharing by encouraging the transfer of university technologies to the private sector. Unfortunately, the present debate over copyright legislation suggests that much of the friction remains. Cost cutting and budget balancing in academic and government funded projects have reduced the output of purely scholarly research making industry the largest producer of scientific data. At the same time, unsure of their present intellectual property rights, many corporate databases are locking down and preventing easy access to their information.

¶2 Science, and especially Bioinformatics, a field that relies on the free flow of data, is hindered by restraints on open access. Though present copyright laws provide some protection in exchange for divulging proprietary information, pending copyright legislation would further empower database owners at the expense

---

[1] Dov Greenbaum is a sixth-year genetics graduate student and Ph.D. candidate in Mark Gerstein's lab at Yale University. He received his bachelor's degree, in Biology and Economics, from Yeshiva University (New York) in 1998. He has written a number of papers on the subjects of genomics, proteomics and databases.

of the academic community. Because the spirit of the law is to promote science, discovery and the growth of the public domain,[3] current legislation must be written with the best interests of Science in mind for "[w]here scientific inquiry is stunted, the intellectual life of the nation dries up."[4]

¶3 This paper, written from the view of a scientist, proceeds as follows: In Section II an overview of Bioinformatics is provided and used to exemplify a science requiring broad access to databases. Section III discusses the problems surrounding database compilations under present copyright law. Section IV draws a historical outline of database legislation, explaining why the database debate has erupted now, and what specific legislative options are on the table. Section V reviews the current debate on pending legislation. Section VI appraises the present situation and shows how database publishers and users are coping without new legislation. Finally, Section VII proposes ideas that should be incorporated into future legislation for a fair and productive solution.

II. WHAT IS BIOINFORMATICS?

¶4 High throughput technologies have transformed biology into a data rich science—labs that were previously examining single genes are now investigating entire genomes.[5] These and other advancements in molecular biology[6] have created a deluge of information that has to be managed.[7] Attempts to find efficient

---

[2] *See* The Bayh-Dole Act, Pub. L. No. 96-517, 94 Stat. 3019 (codified as amended at 35 U.S.C. 200-212 (1980)).
[3] In Graham v. John Deere, 383 U.S. 1, 5 (1966), the Supreme Court reaffirmed the constitutional limits on copyrights and patents ("Congress in the exercise of the patent power may not overreach the restraints imposed by the stated constitutional purpose . . . [and may not recognize exclusive rights] whose effects are to remove existent knowledge from the public domain, or to restrict free access to materials already available. *See also* Yochai Benkler, *SYMPOSIUM: Constitutional Bounds of Database Protection: The Role of Judicial Review in the Creation and Definition of Private Rights in Information*, 15 BERKELEY TECH. L.J. 535 (2000).
[4] ALBERT EINSTEIN, THE WORLD AS I SEE IT (1956).
[5] Traditionally biologists worked on one gene, one protein, or at most, one cellular pathway. Given this narrow research focus, their experimentation did not result in any significant amount of data. The current concept of genome-wide and pan-genomic research was once neither feasible nor practical. The advent of high throughput technologies and computer power has changed this condition, however, and global analyses are now the norm rather than the exception. *See, e.g.* Nicholas Luscombe et al., *What is Bioinformatics? A Proposed Definition and Overview of the Field*, 40 METHODS IN INFORMATICS & MED. 346 (2001).
[6] Many factors came into play at that time. Molecular biology techniques such as the Polymerase Chain Reaction, gene mapping and technological advancements in DNA sequencing were introduced. *See, e.g.*, PAUL RABINOW, MAKING PCR (1996) (discussing polymerase chain reaction); BEVERLY MERTZ PINES, A SHORT HISTORY OF MAPPING: READING THE HUMAN BLUEPRINT (1991) (discussing gene mapping); Walter Gilbert & Alan Maxam, *A New Method for Sequencing DNA*, 74 PROCEEDINGS OF THE NAT'L ACAD. OF SCI. 560 (1977) (discussing advancements in gene sequencing). At the same time, new methods of data analysis as well as technological advances in computer technology, lead to computer scientists joining the ranks of biologists and incorporating algorithms and concepts from computer science in solving many of the new problems such as protein structure and protein sequence.
[7] *See, e.g.,* Almut Schulze & Julian Downward, *Analysis of Gene Expression by Microarrays: Cell Biologist's Gold Mine or Minefield?*, 113 J. CELL SCI. 4151 (2000).

ways to process and administer this data resulted in the field of Bioinformatics, a sort of *in silico*[8] biology. Bioinformatics is:

> Conceptualising biology in terms of molecules (in the sense of physical chemistry) and applying "informatics techniques" (derived from disciplines such as applied maths, computer science and statistics) to understand and organise the information associated with these molecules on a large scale. In short, Bioinformatics is a management information system for molecular biology and has many practical applications.[9]

¶5 Biology now has a plethora of data for this "management information system" to contend with.[10] Given the recent euphoria regarding the sequencing of the human genome, one may be easily convinced that man has conquered nature. However, the media gravely misinforms the public by predicting that science, using the genome, will soon be able to solve all our medical problems. The much-publicised sequencing of the genome is more akin to the Moon Shot. While demonstrating the incredible power of science and technology, it produces little in the way of practical information. The genome may be sequenced, but we do not know what it means. It is up to researchers, and especially those in the field of Bioinformatics, to turn this information into useful and practical results.

¶6 Bioinformatics provides a multi-pronged approach to dealing with, mining, analysing, and deciphering genomic data. Initially, Bioinformatics creates a level of organization allowing researchers to input and access data. Databases are composed of complex architectures designed to integrate radically different forms of data (DNA sequences, images, scientific literature, metabolic pathway data, qualitative protein-protein interaction data, quantitative mRNA and protein expression data, or three dimensional atomic coordinates). Data from multiple databases are extracted, collected and coalesced into more definitive resources, adding value to the information.

¶7 And the genome is but one source of raw data for biological analysis. All life can be divided into layered sets of information. The genome is transcribed by cellular machinery into the Transcriptome. This population of messenger ribonucleic acid (mRNA) is the intermediary between the hard coded DNA of the organism and the proteins that each individual cell produces. Depending on a myriad of influences and factors, each gene may be transcribed into single or multiple copies of mRNA, such that some genes have thousands of copies of their mRNA floating around in a cell. Recently, technologies such as DNA microarrays and oligonucleotide chips have been used to collect this data,[11] although it is still unclear as how

---

[8] *In silico* refers to the concept of doing experimentation solely on a computer. *In silico* is an extension of the concept of *in vitro*, (as opposed to in vivo) where experimentation is done in a test tube.

[9] Our definition of Bioinformatics, submitted to the Oxford English Dictionary. Nicholas M. Luscombe et al., *What is Bioinformatics? An Introduction and Overview*, 2001 INT'L MED. INFORMATICS ASS'N Y.B. 83.

[10] *See, e.g.*, *Drowning in Data*, THE ECONOMIST, June 24, 1999, at 93.

[11] *See, e.g.*, Vivian G. Cheung et al., *Making and Reading Microarrays*, 21 NATURE GENETICS 15 (1999); David J. Duggan et al., *Expression Profiling Using cDNA Microarrays*, 21 NATURE GENETICS 10 (1999).

to effectively process, store and analyse this information.[12] Even less understood is the process of calculating, analysing and comparing it to the Translatome (or Proteome), the population of protein products resulting from the translation of the mRNA.[13]

¶8 Bioinformatics provides methods and algorithms for extracting this data from the databases in a competent fashion. This requires both biological knowledge and computer know-how. BLAST is one such program.[14] Given a sequence, BLAST, a dynamic programming algorithm, runs through a database in search of similar sequences, and outputs a list of such sequences with a statistical significance given for each result. As much of the present functional gene annotation is based upon sequence homology between known and unknown proteins, (i.e. if two genes have similar sequence and one has functional information, while the other does not, the functional information of the former can be applied to the latter) one can appreciate the importance of such an algorithm. BLAST and subsequent similar programs are also designed to take into account evolutionary events that may lead to alternative yet similar sequences.

¶9 Additionally, new tools must be created and other analyses done to produce biologically relevant results. For instance, there is a large body of data resulting from time course experiments: experiments designed to measure mRNA expression using microarrays through a specified course of events in a cell. Each experiment can produce millions of data points. Presently, many researchers attempt to cluster this data using algorithms from varied disciplines.[15] Clustering involves grouping genes (or in this case their mRNA representatives) into sets, wherein each set has the same expression level. The theory goes that those genes that cluster into the same set have similar function.[16] Datamining is also conducted through machine learning techniques.

¶10 Finally, the data has to be presented in a coherent fashion. While this may seem to be a trivial concept, the proper presentation of data, so that complex, mathematically based information can be quickly

---

[12] *See, e.g.*, Mark Gerstein & Ronald Jansen, *The Current Excitement in Bioinformatics-Analysis of Whole-GenomeExpression Data: How Does it Relate to Protein Structure And Function?*, 10 CURRENT OPINION IN STRUCTURAL BIOLOGY 574 (2000).

[13] *See, e.g.*, Dov Greenbaum et al., *Analysis of mRNA Expression And Protein Abundance Data: An Approach for the Comparison of the Enrichment of Features in the Cellular Population of Proteins and Transcripts*, 18 BIOINFORMATICS No. 4, at 585 (2002). *See also* Dov Greenbaum et al., *Interrelating Different Types of Genomic Data, from Proteome to Secretome: Homing in on Function*, 11 GENOME RES. 9 1463 (2001).

[14] *See, e.g.*, Stephen Altschul et al., *Basic Local Alignment Search Tool (BLAST)*, 215 J. MOLECULAR BIOLOGY 403 (1990) (The importance of this algorithm can be recognized in the fact that this paper is the most cited paper in modern biology).

[15] These include, for example, K- means clustering, Self Organizing Maps, and Hidden Markov Models. *See, e.g.*, Arun Jagota, DATA ANALYSIS AND CLASSIFICATION FOR BIOINFORMATICS (2000); Michael. B. Eisen et al., *Cluster Analysis and Display of Genome-Wide Expression Patterns*, 95 PROC. NAT'L ACAD. SCI. U.S.A. 14863 (1998); Laurie. J. Heyer et al., *Exploring Expression Data: Identification and Analysis of Coexpressed Genes*, 9 GENOME RES. 1106-15 (1999).

[16] *See, e.g.*, Jiang Qian et al., *Beyond Synexpression Relationships: Local Clustering of Time-shifted and Inverted Gene Expression Profiles Identifies New, Biologically Relevant Interactions*, 314 J. MOLECULAR BIOLOGY 1053 (2001).

and easily understood by other researchers, is fundamental and non-trivial.[17] In many cases, a researcher will integrate many large, disparate and redundant datasets to form an extended and complete database of all the pertinent data.[18]

¶11 Given the operation of Bioinformatics research, overall it should be obvious that the field requires unhindered access to databases in order to survive—databases are the raw material of Bioinformatics. Consequently, significant limitations imposed by Congress or database owners will likely be detrimental to the field and scientific discovery more generally.

III. DATABASES—THE LEGAL PROBLEM

¶12 While printed compilations have always been protected under copyright law, the protection of computer databases is fairly recent.[19] As with all copyright law, copyright on databases protects only original works. As such, in most instances only the layout the database is protected and not the inherent data itself. Prior to *Feist Publications v. Rural Telephone Service*,[20] the general practice was for the courts to allow for database owners to argue that their efforts—by "sweat of the brow"—constituted reason enough to protect a database and its data under copyright.[21] In *Feist*, however, the Supreme Court held that collections of facts could be granted only "thin" protection, that is, only the arrangement of the database would be protected. The underlying data would be part of the public domain if not novel[22] as only originators may receive the benefits of copyright—"Only those who add to human knowledge may receive an exclusive right in what they added."[23]

¶13 The problem for many scientific fields and the databases that service these fields is that they *only* deal with compilations of fact.[24] *Feist* stated that "facts do not emerge from scientists' efforts, but rather

---

[17] Edward Tufte, Professor Emeritus of Political Science, Computer Science & Statistics & Sr. Critic Graphic Design, Yale University Graduate School's Tercentennial lecture series In the Company of Scholars, Nov. 15, 2000.

[18] *See, e.g.*, Degen Zhuo et al., *Assembly, Annotation, and Integration of UNIGENE Clusters into the HumanGenomeDraft*, 11 GENOME RES. 904 (2001).

[19] The National Commission of New Technological Uses of Copyrighted Works (CONTU) determined in 1978 that computer databases should fall under the protection of the copyright act. *See* Jennifer Askanazi et al., *The Future of Database Protection in US Copyright Law*, 2001 DUKE L. & TECH. REV. 17, ¶5.

[20] 499 U.S. 340 (1991).

[21] *See, e.g.*, Southern Bell Tel. & Tel. v. Associated Telephone Directory Publishers, 756 F.2d 801 (11th Cir. 1985); West Publishing Co. v. Mead Data Central, 616 F. Supp. 1571 (D. Minn. 1985).

[22] Copyright law does not prohibit the copying of facts, even newly discovered or expensively acquired facts. *See* Harper & Row Publishers, Inc. v. Nation Enterprises, 471 U.S. 539, 556 (1985); Kregos v. Associated Press, 937 F.2d 700, 703-10 (2d Cir. 1991).

[23] *See, e.g.*, Benkler, *supra* note 3, at 545.

[24] This is contrasted with the postmodern view that scientific research involves a "de minimis of quantum creativity," that the scientists' data are the results of the human social and political situation, and as such are not an objective truth. This is ridiculed in the scientific community. The reality lies somewhere in the middle. Data, in essence, does not attain the status of fact until it is peer reviewed or even replicated. Moreover, many times that which is published is not the raw data, but rather data that has been 'massaged' to account for irregularities. Taken as a whole though, it can be assumed for the most part that data is not novel. *See, e.g.*, CORYNNE MCSHERRY, WHO

they copy them from the world around them. Raw facts may be copied at will. This result is neither unfair nor unfortunate. It is the means by which copyright advances the progress of science."[25] Similarly, Nimmer claims that "[t]he discoverer of a scientific fact . . . may not claim to be the author of the fact."[26] Additionally, courts have differentiated between facts that are "hard ideas" and should remain in the public domain and "soft ideas" which are "infused with the author's taste or opinion."[27]

¶14 These distinctions are quite imprecise, creating a general air of confusion. Database owners, unsure of their rights, create long and complicated licenses in an effort to protect their investments from competitors.[28] Academia, also unsure of its rights, counters with long and complex negotiations, to insure that it is not being roped into an unfair situation. In addition, many owners of scientifically important databases recoil into the absolute protection of trade secret laws, further hindering scientific research.[29]

¶15 A case in point: determining qualitative protein expression is a very laborious, expensive, and tedious process.[30] This has drastically limited the number of experiments conducted in academia and the overall data available for scholarship. However, many trans-national pharmaceutical conglomerates have the resources to conduct these experiments and have databases overflowing with this information.[31] Yet, because of the above-mentioned uncertainties and fears they do not share their information.

¶16 Some commentators advocate the protection of databases through licensing and other "self help" methods. The Digital Millenium Copyright Act that prevents the circumvention of these technological locks further strengthens these methods, and, in some cases, prevents access even for fair use applications.[32]

OWNS ACADEMIC WORK (2001); Roger Kimball, *A Painful Sting Within the Academic Hive*, THE WALL ST. J., May 29, 1996, at A18.

[25] Robert Gibbons, *Database Protection*, 11 FORDHAM INTELL. PROP., MEDIA & ENT. L.J. 275, 287 (2001) (quoting Feist Publ'ns v. Rural Tel. Serv. Co., 499 U.S. 340, 349 (1991)).

[26] MelvilleB. Nimmer, *The Subject Matter of Copyright Under the Act of 1976*, 24 UCLA L. REV. 978, 1001-1003 (1977).

[27] *See* CCC Info. Servs., Inc. v. Maclean Hunter Mkt. Reports, 44 F.3d 61, 72 (1994).

[28] THE NAT'L ACAD. OF SCI., PROCEEDINGS OF THE WORKSHOP ON PROMOTING ACCESS TO SCIENTIFIC AND TECHNICAL DATA FOR THE PUBLIC INTEREST: AN ASSESSMENT OF POLICY OPTIONS §§ 7-8 (1999), *available at* http://www.nap.edu/html/proceedings_sci_tech/.

[29] Personal communication with researchers in private pharmaceutical industries, although most lawyers employed by these companies are unwilling to talk about this situation.

[30] *See, e.g.*, Julio E Celis and Pavel Gromov, *2D protein Electrophoresis: Can it be Perfected?*, 10 CURRENT OPINION IN BIOTECH. 16 (1999).

[31] As reported to the author by researchers at multiple companies.

[32] *See, e.g.*, Jeff Sharp, *Coming Soon To Pay-Per-View: How The Digital Millennium Copyright Act Enables Digital Content Owners To Circumvent Educational Fair Use*, 40 AM. BUS. L.J. 1 (2002).

¶17 While some database providers license out their data, it is not uncommon for a license to be tens if not hundreds of thousands of dollars, and governmental funding agencies, which provide most of the grant money for scientific research, are notoriously reluctant to provide extra funds for these licences.[33]

¶18 It is unclear whether additional protection would increase the sharing of data, or lock up more data in the private domain. Would it provide incentives for additional investment in database production? Some studies show that there would be no effect.[34] What is needed is a uniform federal law that, while not expanding copyright protection to scientific fact, fairly protects the investments of the database owner and also promotes the advancement of science.

IV. DATABASES–LEGISLATIVE HISTORY

¶19 In 1996 the European Union issued a directive[35] that gave considerable more protection to databases than provided by present American law.[36] This directive afforded two tiers of protection: (1) copyright protection for the aesthetics of the database, i.e. selection and arrangement, and (2) protection against the unfair extraction of significant portions of the data.[37] This second sui generis protection affords the database owner fifteen years of protection, which is given each and every time the database changes (i.e. an update); in essence, protection in perpetuity.[38] While the EU directive's implicit goal was for harmonization of the member states' copyright laws, and thus required that all eighteen countries institute the policies by January 1998, to date only thirteen countries have implemented it.[39] In addition, the Directive afforded an escape clause requiring a follow-up in 2001, at which point, if the sui generis database protection was responsible for decreasing competition, a non-voluntary license requirement might be imposed.[40]

¶20 European courts have construed the directive to grant excessive protection to database owners. While the language of the directive requires that the database be arranged in a systematic way, the

---

[33] NATIONAL RES. COUNCIL, BITS OF POWER—ISSUES IN GLOBAL ACCESS TO SCIENTIFIC DATA (1997) (Referencing statements of R. Stephen Berry, Chair, professor of chemistry and the James Franck Distinguished Service Professor at the University of Chicago) *available at* http://search.nap.edu/readingroom/books/BitsOfPower/.

[34] *See, e.g.*, NAT'L RES. COUNCIL, WORKSHOP ON PROMOTING ACCESS TO SCIENTIFIC AND TECHNICAL DATA FOR THE PUBLIC INTEREST (1999), *available at* http://www.nap.edu/html/proceedings_sci_tech/.

[35] Council Directive 96/9/EC of 11 March 1996 on the legal protection of databases, 1996 OJ (L 077).

[36] This directive was based on the "catalog rule" of some Scandinavian countries, which granted non-copyrightable compilations some short-term protection against reproduction. *See, e.g.*, Mark Schneider, *The European Union Database Directive*, 13 BERKELEY TECH. L.J. 551, 555 (1998)

[37] Council Directive 96/9/EC of 11 March 1996 on the legal protection of databases, art. 7-8, 1996 OJ (L 077)

[38] *See Id.* at art. 10.

[39] P. BERNT HUGENHOLTZ, THE NEW DATABASE RIGHT: EARLY CASE LAW FROM EUROPE (2001), *at* http://www.ivir.nl/publications/hugenholtz/fordham2001.html.

[40] GREGORY AHARONIAN, INFORMAL NOTES: THE FUTURE OF INTELLECTUAL PROPERTY ON THE INTERNET: A DEBATE (2000), *available at* http://cyber.law.harvard.edu/futureofip/notes.asp.

present situation allows for any random or even trivial selections of data to be protected.[41] Additionally, the concept of a substantial investment (a requirement for protection) as well as many other terms are weak and poorly defined.[42] Most important for academia, the EU directive does not explicitly grant exemption status to anyone,[43] including research institutions, although it allows for individual countries to do so. France is the only country so far to not allow any exemptions.[44] Already, there are complaints in the scientific community that these statutes are making European collaborators reluctant to share.[45] Most important for the database industry, the directive only protects databases residing in countries that institute the directive.[46] As such, American databases are left at a significant disadvantage in Europe.

¶21 There have been many attempts to beef-up US copyright laws to protect databases (partially as a result of the EU's directive). H.R. 3531, modelled after the EU Directive, was introduced in 1996.[47] The bill provided no exceptions for fair use, granted twenty-five years (ten more than the EU) of protection, and had severe criminal penalties.[48] Alarmed by the extent of the law,[49] scientists rallied against it in a letter writing campaign to Rep. Moorhead, sponsor of the bill. Taken aback by the opposition, the house never brought the bill up to vote.[50] The following year H.R. 2652 was introduced.[51] This bill, while more favourable to science, was unacceptable. For instance, it did not allow for fair use of the copyrighted material if a potential market could be harmed. The bill was eventually tabled.[52] There were two bills in the 106th Congress and both are still on the table: The Collections of Information Antipiracy

---

[41] "(2) For the purposes of this Directive, 'database' shall mean a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means." *See* Council Directive 96/9/EC of 11 March 1996 on the legal protection of databases, art. 1, 1996 OJ (L 077).

[42] STEPHEN MAURER, RAW KNOWLEDGE: PROTECTING TECHNICAL DATABASES FOR SCIENCE AND INDUSTRY 366 (1999), *available at* http://books.nap.edu/books/NI000903/html/337.html.

[43] Exemption status means that academics or educators would not be liable in a lawsuit, as long as the copyright violation was conducted for educational or research purposes.

[44] ANNE LINN, HISTORY OF DATABASE PROTECTION: LEGAL ISSUES OF CONCERN TO THE SCIENTIFIC COMMUNITY 5 (2000), *available at* http://globalchange.gov/policies/linn.html.

[45] Stephen Maurer & Suzanne Scotchmer, *Database Protection: Is It Broken and Should We Fix It?*, 284 SCIENCE 1129 (1999).

[46] *See* Council Directive 96/9/EC of 11 March 1996 on the legal protection of databases, stat. 56, 1996 OJ (L 077).

[47] *See* H.R. 3531, 104th Cong. (1996).

[48] *See Id.* at §§ 6, 8.

[49] "Provisions in the bill represent an unprecedented erosion of the rights of the American research and education community and would have detrimental impacts on the conduct of research and education activities in this country." Peter Jaszi, *Some Public Interest Considerations Relating to H.R. 3531 Database Investment and Intellectual Property Antipiracy Act of 1996*, FED. RELATIONS E-NEWS (Aug. 1996), *at* http://arl.cni.org/info/frn/copy/peter.html.

[50] Jonathan Band & Jonathan Gowdy, *Sui Generis Database Protection Has its Time Come?*, D-LIB MAGAZINE (1997), *at* http://www.dlib.org/dlib/june97/06band.html.

[51] H.R. 2652, 105th Cong. (1998).

[52] LINN, *supra* note 44, at 7-8.

Act[53] and the Consumer and Investors Access to Information Act.[54] While both bills attempt to protect databases, they employ different methodologies.

¶22 Both Bills are problematic. Congress seems to be reacting to a general sense that data pirates are trying to steal information from databases. The lack of fact finding regarding the "database industry, its economic basis, its conditions for further growth is quite stringent in the congressional record . . . the debate is premised on presupposition rather than fact . . . each bill lacks an empirical basis that justifies its enclosure of significant parts of the public domain."[55] Congress indicated this discomfort with both bills by including a reporting requirement thirty-six months after inception, to prove that the legislation is both necessary and good for the market.[56]

¶23 H.R. 354 is similar to the E.U. directive, as it provides *sui generis* protection to databases in the form of ownership of the data in the database, and protects many databases not currently covered under present copyright law.[57] The bill creates broad prohibitions and then carves out vague exemptions, such as the exemption from liability for using an insubstantial portion of the collection of data though insubstantial was never defined. Conversely, H.R. 1858 delineates narrow prohibitions, and allows transformative uses of the data,[58] which are not allowed under H.R. 354.[59] H.R. 354 contains very narrow fair use exemptions, and does not exempt research that may be part of or provide the foundation for an actual or potential market. Additionally, it places the burden of proof, i.e. that harm was not caused, on the user.[60] In contrast, H.R. 1858 allows any extraction of data for scientific, research and educational purposes so long as it does not cause substantial harm to the database owners.[61]

¶24 Additionally, there is concern that H.R. 1858 may fail to provide enough protection to avail U.S. databases of the reciprocal protection provisions under the European Union's directive.[62] Furthermore,

---

[53] H.R. 354, 106th Cong. (1999).
[54] H.R. 1858**,** 106th Cong. (1999).
[55] Marci Hamilton, *Symposium: A Response to Professor Benkler*, 15 BERKELEY TECH. L.J. 605, 626 (2000).
[56] *Id.*
[57] Congress has been careful when contemplating new legislation so as to not extend copyright protection. Both of the present bills expressly state that they are not copyright bills. H.R. 1858**,** 106th Cong. §105 (1999); H.R. 354, 106th Cong. §1405 (1999). In fact, 1858 goes to such an extent to distance itself from copyright as to not put on a time limit on the protection. *See* H.R. 1858**,** 106th Cong. (1999).
[58] *See Id* at §102.
[59] *See* H.R. 354, 106th Cong. §1403 (1999).
[60] Michael Freno, *Database Protection: Resolving the U.S. Database Dilemma with an Eye Toward International Protection*, 34 CORNELL INT'L L.J. 165, 202-03 (2001).
[61] *See* H.R. 1858**,** 106th Cong. §103 (1999).
[62] "Whereas the right to prevent unauthorized extraction and/or re-utilization in respect of a database should apply to databases whose makers are nationals or habitual residents of third countries or to those produced by legal persons not established in a Member State, within the meaning of the Treaty, *only if such third countries offer comparable protection to databases produced by nationals of a Member State or persons who have their habitual residence in the territory of the Community*" (emphasis added). *See* Council Directive 96/9/EC of 11 March 1996 on the legal

there are many who feel that this bill is too narrow to provide enough protection for the domestic database market.[63] H.R. 1858 is enforceable only by the FTC, which has little experience in intellectual property or the necessary enforcement clout.[64] H.R. 354 is further flawed because it provides sweeping one-sided protection only to the database industry. The text of the bill is vague, and as such, can and would be abused by database owners at the expense of the public domain.

V. DATABASES–THE DEBATE

¶25 Given the strong intellectual property protection championed by H.R. 354, it is supported by companies such as Reed-Elsevier and eBay.[65] However, most of the scientific community, including the National Academies of Science,[66] supports the rival H.R. 1858 which provides researchers with broader fair use exemptions.[67] This last group also includes Yahoo, America Online and Charles Schwab:[68]

¶26 Opponents of new far-reaching legislation claim that:

- The increase in encryption and password protection on the web,[69] coupled with the need to constantly update databases in real-time, create a difficult situation for would-be database pirates–obviating the need for legislative limitations;

- Many databases, like brand-name products, create a sense of trust and loyalty among their users, who would not turn to pirates. Law firms, for example, would not use pirated data from a legal database, as they would not be able to rely on the accuracy of the information;

---

protection of databases, stat. 56, 1996 OJ (L 077). As such, HR 1858 whose detractors claim is not as strong as the EU directive, would put American databases at risk.

[63] Gail Hodge, *After the EU Directive: The Impact of European Database Protection on Stakeholder Groups*, 26 BULL. OF THE AM. SOC. FOR INFORM. SCI., (Feb./Mar. 2000).

[64] David Carney, *Summary of Bills Pertaining to Protecting Databases from Piracy in the 106th Congress*, TECH L.J. (Aug. 1999), *at* http://www.techlawjournal.com/cong106/database/Default.htm (last visited May 29, 2003).

[65] For a discussion of support surrounding H.R. 354 and H.R. 1858 *see* David Baron, eBay and Database Protection, Case Study No. P-33, Graduate School of Business, Stanford University, Sept. 2000, *available at* http://www.gsb.stanford.edu/cebc/pdfs/C_P_033_case.pdf.

[66] Joshua Lederberg, a Nobel Prize Laureate and past-president of Rockefeller University, put forth many of his reasons for opposing H.R. 354 in his testimony before Congress, including (i) that the bill would reduce competition (ii) that 354 defines databases to broadly, potentially including any information including literary works (iii) that the vagueness of the bill creates uncertainties in the law (iv) the length of protection could extend in perpetuity (v) the lack of public interest exceptions (vi) the lack of requirements for reasonable licensing terms. (vii) the strong civil and criminal penalties the bill creates will prevent the open exchange of data for fear of retribution. *See* Joshua Lederberg, *Statement on behalf of the National Academy of Sciences National Academy of Engineering Institute of Medicine and American Association for the Advancement of Science before the Committee on the Judiciary U.S. House of Representatives*, (Mar. 8, 1999) (transcript *available at* http://www.arl.org/info/letters/lederbergtest.html).

[67] Jocelyn Kaiser, *Scientists Decry Antipiracy Bill*, 286 SCIENCE 1568 (1999).

[68] *See* Baron, *supra* note 65.

[69] *See, e.g.*, LAWRENCE LESSIG, CODE AND OTHER LAWS OF CYBERSPACE (1999) (claiming that, because of technology, copyrights are more effectively protected than at any time since Gutenberg).

- While the database publishers claim that protection provides incentives for further database publishing, they are in effect "over fencing"–erecting too many artificial barriers to the production of science;[70] and

- Increased legislation and protection will lead to monopolies, as only a few providers service many of the markets.[71]

¶27 On the other side of the debate, those in favour of strong legislation claim:

- That businesses must expend significant time, effort and money to create quality databases, while those who pirate and free load do so with minimal effort;

- Broader definitions of violations, fewer exceptions, and greater penalties are required to prevent extensive database piracy and the eventual collapse of the database market;

- Present requirements for copyright protection would prevent many compilations from being copyrighted, as they are essentially compilations of pure fact;

- Licensing databases necessitates cumbersome complications in database usage, requiring manufactures to double check to confirm that they do not violate any licenses. Database providers find that licenses are tedious to negotiate, difficult to enforce, and that they are limited to the parties in the contract and not any downstream users of the database. Additionally the courts have refused enforcement of contracts that have been "unconscionable" or "contrary to public policies relating to innovation, competition and free expression."[72]

VI. DATABASES–THE PRESENT

¶28 This is the future the proponents of H.R. 354 want you to believe in: the once strong and vibrant database community is falling apart. America is losing its foothold in the information industry. As with the music and entertainment industry before it, the database industry is now the target of hackers who disregard lax and unenforceable copyright laws. Consumers are unwilling to pay original developers the fair market value for the data when they can get it cheaper elsewhere. Those few American companies that still have valuable information are doing their best to keep their data a trade secret. This reduces the amount of

---

[70] *See, e.g.*, Paul David, *Tragedy of the Public Knowledge 'Commons'? Global Science, Intellectual Property and the Digital Technology Boomerang*, ELECTRONIC J. INTELL. PROP. RIGHTS (2001), *at* http://www.oiprc.ox.ac.uk/EJWP0400.pdf.

[71] J.H. Reichman, Professor of Law, Duke University School of Law, Statement concerning H.R. 2652 before the Subcommittee on Courts and Intellectual Property Committee on the Judiciary House of Representatives (October 23, 1997) (transcript *available at* http://www.house.gov/judiciary/41121.htm).

[72] *See, e.g.*, MAURER *supra* note 42.

information in the public domain, and impedes the growth of science and technology in the US. Meanwhile, the European databases, under strong protection, are growing and profitable.

¶29 In the immediate wake of the European Union's Directive, Congress almost believed this. Nothing could be further from the truth. Independent of perceived threats, the American database industry is thriving.[73] Most databases use "self help" methods to protect their investments, independent of limiting copyright legislation. These include (1) bilateral contracts with restrictive licensing agreements[74] (2) "shrinkwrap/clickwrap" protection[75] (3) bundling with other copyrightable material (4) continuous updating, thus requiring a pirate to sink significant costs into any pirating endeavour (5) allowing only partial access through restrictive websites (6) and passwords or encryption.[76] Other legal alternatives to copyright legislation include trade secret,[77] criminal, and tort laws.[78]

VII. POSSIBLE SOLUTIONS

¶30 Congress should commission a serious bipartisan (i.e. both publishers and end-users) fact finding commission into the actual state of affairs in the database industry and academic institutions. This would be in contrast to previous copyright issues where only the narrowly interested copyright lawyers hammered out

---

[73] The American Database industry seems to be booming, and it does not seem to be worried about *Feist* or the European Union. *See* J. Saltiel, Note, *With Nowhere Else to Hide Can the First Amendment Protect Databases*, 2001 U. ILL. J.L. TECH. & POL'Y 163, 186 (2001). Since *Feist* the number of databases in the United States has grow at a phenomenal rate. From 1991 to 1997 the number of database increased 35% from 7637 to 10338. This increase accurately reflects an investment of the private sector as during the same period, privately funded databases increased their market share by 8%, that is by 1997 78% of all databases in the US were privately owned. ASS'N OF RES. LIBRARIES, MYTHS AND FACTS ABOUT S. 2291/TITLE V OF H.R. 2281 (USED TO BE H.R. 2652): THE COLLECTIONS OF INFORMATION ANTIPIRACY ACT (1998), *at* http://www.arl.org/info/frn/copy/myth.html.

[74] Many companies polled claimed that this was their initial line of defense. *See* NAT'L RES. COUNCIL, COMMERCIAL SECTOR DATA: PROCEEDINGS OF THE WORKSHOP ON PROMOTING ACCESS TO SCIENTIFIC AND TECHNICAL DATA FOR THE PUBLIC INTEREST (1999), *available at* http://books.nap.edu/books/NI000903/html/R1.html.

[75] Many purveyors of digital databases employ the shrinkwrap/clickwrap protection. Essentially, the user agrees to licensing requirements prior to opening the shrinkwrap on software or clicking on an agreement on the Internet. Shrinkwrap licenses have been upheld in ProCD Inc v Zeidenberg, 86 F.3d 1447 (7th Cir. 1996); Mortenson v. Timberline Software Corp., 140 Wn. 2d. 586 (Wash. 2000); Hotmail Corporation v. Van$ Money Pie, 1998 U.S. Dist. LEXIS 10729 (N.D. Cal. 1998). Several courts, however, have not enforced shrinkwrap agreements. *See, e.g.*, Ticketmaster Corp. v. Tickets.com, 2000 U.S. Dist. LEXIS 4553 (C.D. Cal. 2000) (license not sufficiently visible); Vault Court v Quaid Software Ltd, 847 F.2d 255 (5th Cir. 1988) (invalidating a contraction restriction against decompiling the software). *See also* MAURER, *supra* note 42; Garry Founds, *Shrinkwrap and Clickwrap agreements: 2b or not 2B*, 52 FED. COMM. L.J. 99 (1999).

[76] *See* MAURER, *supra* note 42.

[77] While companies have protected their databases through trade secret litigation, *see, e.g.*, Florida Power & Light Co. v. Util. Serv. of Am., Inc., 550 So. 2d 13 (Fla. Dist. Ct. App. 1989), companies cannot claim trade secret rights over all their data, as this invalidates the protection. Ralph Losey, PRACTICAL AND LEGAL PROTECTION OF COMPUTER DATABASES, *at* http://floridalawfirm.com/article.html.

[78] Possible tort actions could include trespass, conversion and theft. However, in these cases it would be difficult for database owners to be fully compensated for their loss. Moreover, compensation would vary from state to state. *See* Saltiel, *supra* note 73, at 170-01.

legislation for their politically astute clients.[79] The new database bill should strive to include the broadest audience in the bill drafting process. New legislation should also take into account that there are many unique markets for databases. Moreover, any new policy must recognize that the ultimate purpose of copyright is to provide incentives for creativity and the sharing of that research "[t]o promote the Progress of Science and useful Arts"–not simply to profit industry. Below, I offer some specific suggestions to fulfil these general objectives.

¶31 First, the legislation should be as clear and concise as possible. Clarity includes the need to ensure that: (i) databases should be defined so that the term does not include frivolous collections of data or journal articles; (ii) scientific research should have broad fair use exemptions; and (iii) and scientific fact should not be included in any protection regime.

¶32 Second, the courts should take into account the intent of an infringer and the actual harm the infringer causes. Database publishers should only be able to claim "harm" in their primary market, excluding secondary and future markets. The concept of qualitative harm is amorphous, and, as such, the concept of harm should be limited to only quantitative extraction of data. While malicious intent should be prohibited and penalties be imposed, courts should be lenient when there is no malice. The burden of proof should fall on the copyright owner to show that he has been harmed. Most scientific research would thus be protected from lawsuits.

¶33 Third, protection should only be for a limited length of time, consistent with the rate of change in the general field. Present legislation's limit of fifteen years is too long for data that is constantly changing. Upgrades should not extend the life of the protection.

¶34 Fourth, legislation should require that licenses be structured to take into account the cultural norms of the scientific community, including pervasive sharing and publishing of data. This should include: (i) Encouraging an industry wide licensing standard because many scientists do not have the know-how to negotiate licenses both for the acquisition and the dissemination of data, the process of licensing is inefficient for science[80] and provides disincentives for industry-academia partnerships;[81] (ii) Industry should be encouraged, perhaps with government incentives, to license data to academic institutions at only marginal cost;[82] (iii) Data transformed through creative means should become exempt from original licensing restrictions and allowed to be shared freely; (iv) Licenses should not be allowed to overturn any legislated exemptions.

---

[79] *See, e.g.*, JESSICA LITMAN, DIGITAL COPYRIGHT: PROTECTING INTELLECTUAL PROPERTY ON THE INTERNET (2001).
[80] *See, e.g.*, MYRA WILLIAMS, PROCEEDINGS OF THE WORKSHOP ON PROMOTING ACCESS TO SCIENTIFIC AND TECHNICAL DATA FOR THE PUBLIC INTEREST (1999), *available at* http://books.nap.edu/books/NI000903/html/R1.html.
[81] Personal communication with an administrator at a biotech company.
[82] *See* J.H. Reichman & Paul Uhlir, *Promoting Public Good Uses of Scientific Data A Contractually Reconstructed Commons for Science and Innovation*, 66 LAW & CONTEMP. PROBS. 315, 329 (Spr. 2003).

¶35 Fifth, the DMCA should be changed to provide for fair use access to databases. As it stands, a database provider can protect non-copyrightable material from being extracted using the DMCA. Liability type rules similar to that of fair use should be incorporated into this act.

¶36 Economic solutions such as price discrimination wherein academic users would have access to data at marginal cost (in the digital world marginal cost is very low) or time embargoes are also available to database producers. Recently some prominent scientific journals (i.e. Nature) have experimented with limiting access for libraries with site licenses when the journals are the most commercially valuable (during the first three months) and then following this black out period, allowing them to be broadly accessed. This concept is similar to a movie studio releasing less profitable videocassettes for rental only after the movie has been shown for a certain amount of time in the more profitable theatres.

¶37 Finally, and perhaps most importantly, lawmakers should realize that the economics of databases remains unclear. As such, any law should include a sunset provision so that the costs and benefits of any legislation can be properly evaluated.

VIII. CONCLUSION

> "While I shall think myself bound to secure every man in the enjoyment of his copyright, one must not put manacles upon science."[83]

¶38 Any legislation affecting copyright must be written, first and foremost, with the best interests of science in mind. It should mimic science in its openness and rigorousness. Those few limitations that are placed on the usage of databases, restricting those in the service of science and the public domain, should be clear, forthright and unambiguous. Science's most steadfast law, Occam's Razor, claims that the simplest route is the best.

[83] Carey v. Kearsley, 4 Esp. 168, 170, 170 Eng. Rep. 679, 681 (K.B. 1803).