A universal legal framework as a prerequisite for database interoperability

Dov Greenbaum & Mark Gerstein

Databases are fundamental to modern scientific research, both as archives and, via manipulation of their contents, as research tools in their own right. One obvious example is the annotation of genomes, requiring systematic downloading, reformatting, standardizing and combining of data in a unified computational framework. This process requires both repeated access to databases, and the ability to show the transformed data, repackaged in a new format, alongside the evidence—the original data sets.

It is obvious that interoperation of databases through universal scientific formats and standards facilitates research; data are ineffectual if scattered among incompatible resources. Not as obvious is the need for robust legal frameworks to ensure interoperation. The ambiguity of the present copyright laws governing the protection of databases creates a situation where researchers are unclear about their rights to extract and combine data; and database owners, unsure of how laws safeguard their information, overprotect their data with licenses and technological mechanisms that impede interoperation.

Much of the current international database debate can be described as responsive volleys of legislation across the Atlantic, each side trying to establish an industry-wide level of protection. Thus, responding to judicial and European developments in database protection, the US Congress (Washington, DC, USA) is currently attempting to augment weak copyright protections. In doing so, US lawmakers need to consider the repercussions of their legislation on scientific research.

There is no doubt that database protection is necessary. However, science advances through building upon previous research. Thus, scientific researchers (both academic and commercial) who depend on access to these databases require legislation that is narrow in scope and broad in academic exemptions, and that encourages data sharing and limits the application of technological safeguards that inhibit interoperability. By creating a system that limits the ability of database owners to incorporate technical safeguards and yet offers substantial legal protections under a compulsory license scheme, legislators can help create a universal standard of protection that is favorable to scientific research.

This directive has the effect of separating EU scientists from their international counterparts, limiting collaboration.

Legal history

The present situation has resulted from a confluence of recent court rulings and legislative actions (see **Table 1**). US copyright law has generally tried to balance the constitutional imperative of promoting the 'progress of science and useful arts' with the need to provide incentives to authors and producers of original works; this was typically realized through granting monopolies limited in their duration, power and scope.

Databases have always held a precarious position within copyright. Initially, the general perception was that the efforts of the database author ('sweat of the brow') fulfilled a threshold requirement for protection. Still, although the data were protected from unjustifiable infringement, the law provided for 'fair use' reutilization of the data, for example, in academic research.

In a 1991 landmark decision (*Feist v. Rural*)[1], the US Supreme Court questioned the copyright protections granted to factual databases; the court argued that the "*sine qua non* of copyright is originality" and as such the "discoverer of a scientific fact…may not claim to be the author of the fact….The discoverer merely finds and records." So although the architecture and interface of a database may be original and protectable, the factual data within are not.

In 1996, the European Commission (EC; Brussels, Belgium) instituted a Database Directive (96/9/EC) granting databases considerable protection over and above previous international norms, and provided these rights only to reciprocating nations (practically, other European Union (EU) countries), giving them a significant advantage over foreign competitors whose legislators would not counter with equally tough protections. This directive has the effect of separating EU scientists from their international counterparts, limiting collaboration. Some databases already explicitly cite these EU protections in limiting data extraction (*e.g.*, the Human Gene Mutation Database).

Moreover, the directive may also impede academic research within the EU. The UK Royal Society (London, UK) has highlighted some of these problems; although there are 'fair-use' exceptions for academic use, the directive does not require EU member countries to implement them in their

Dov Greenbaum is in the Department of Genetics and Mark Gerstein is in the Departments of Genetics, Molecular Biophysics & Biochemistry, and Computer Science, Yale University, 266 Whitney Avenue, PO Box 208114, New Haven, Connecticut 06520, USA. e-mail: dov.greenbaum@yale.edu

own legislation. Additionally, fair-use exemptions are only permitted with regard to data extraction, not reuse.

Finally, databases are allowed a new term of protection for every update, in essence granting copyright protection in perpetuity, greatly expanding the scope of protection and preventing data from ever falling into the public domain. Imagine the *Principia Mathematica* never being freely available.

In a knee-jerk reaction, the US Database Investment and Intellectual Property Antipiracy Act (HR 3531) was drafted, granting even more protection to databases, but it died in the House. Subsequent bills, most supported by little if any empirical evidence of necessity, have been introduced, but none have passed through Congress. Ongoing closed-door sessions could result in a compromise bill in the near future.

Concurrently, US courts have limited fair-use exemptions. Clickwrap licenses (*e.g.*, pop-up windows asking the user to agree to the terms of a software application) were ruled sufficient to limit the user to a contracted agreement (*ProCD v. Zeidenberg*)[2], even though some generic user rights could be rescinded through these agreements. By allowing nonnegotiable and inconsistent contracts to control access to databases, the US courts have

Table 1 A summary of recent major legislation in the United States and Europe and its effects on scientific research

			Previously proposed US legislation	
	EC directive	US Digital Millennium Copyright Act	H.R. 354	H.R. 1858
Description	A *sui generis* protection limiting the extraction and/or reuse of data from databases.	Disallows circumvention of protective technological measures used by copyright owners.	*Sui generis* protection to databases and their data not covered under present copyright law.	Targets commercial pirates but does not prohibit the transformative use of data.
Protects	Database content. "[A] substantial part evaluated qualitatively and/or quantitatively, of the contents of the database."	Any copyrighted work. Data in databases is protected because access to uncopyrightable data is tied to access to the copyrighted portions of the database.	Database content. Prevents the extraction of substantial (quantitative and/or qualitative) part of a database that has been created through an outlay of time even for transformative use.	Similar to H.R. 354, except that it only protects databeses in cases where a duplicate (a concept that will be defined by the courts) has been made without permission.
Exemptions	Narrow and optional. Allowing for fair use "where it is use for the sole purpose of illustration for teaching and scientific research."	Narrow. Prohibits even non-infringing uses.	Narrow. Academic use is permitted so long as it does not harm primary or future markets.	Broad. For all scientific work "so long as such conduct is not part of a consistent pattern engaged in for the purpose of direct commercial competition."
Penalties	Differs among EU countries.	Civil & criminal.	Civil & criminal.	Civil. Administered by the US Federal Trade Commission.
Term	Potentially perpetual. A database is given a new 15-year term after every upgrade.	Perpetual	15 years	Perpetual. The bill is based on misappropriation, not intellectual property.
Relevance to research	1. Limits international collaboration. The EC directive has a reciprocity clause. Only databases from countries with similar levels of protection will will be provided *sui generis* protection in Europe. 2. Ambiguous. Users may not know what is a 'qualitively or quantitatively significant amount of data' to trigger infringements. 3. Potential issues with data reutilization and transformation. Fair use, when applied, does not allow for reutilization or transformative use of the data.	Limits access to any copyrighted work that is stored digitally. Provides further protection to 'unprotectable' facts, potentially impeding their usage by the scientific community.	1. Ambiguous. May limit the extraction of data because of uncertainties incorporated into the law (that is, a user may not be able to be to readily determine whether they are extracting a "qualitatively or quantitatively large portion" of a database or if the extraction will harm a future market). 2. Fear of criminal prosecution may also serve as a disincentive to extract data.	1. Enforcement issues may encourage digital safeguards. Given that penalties are under control of the overburdened FTC, owners may feel as if they are not significantly protected and thus resort to restrictive technological safeguards. 2. Exemptions are still not broad enough for bioinformatics research. Although generally favorable to scientific research, it may not allow extraction of entire data sets from databases for whole-genome research. Additionally, data is protected indefinitely, never moving into the public domain.

effectively empowered sole source providers of irreproducible data to charge licensing fees under limiting conditions.

One unifying theme in all the database legislation to date has been the absence of a clear definition of a database. Colloquially, databases are defined as organized, indexed collections that allow users to efficiently access and organize heterogeneous information. Internally, many databases have complex tabular structures organized by a specialized database management system, which provide a bridge between the raw data and the end user. However, in the various attempts to legislate databases, the term has been defined too broadly. In Europe, especially, this has led to a situation where the courts, interpreting the law, have extended protection to even trivial lists of facts[3].

Technological safeguards

Even without a clear legal structure for protection, the database industry is growing[4] and almost all the major vendors plan on launching new products this year[5]. Meanwhile, the legal uncertainty has resulted in an explosion of technological safeguards, far more limiting than any law in their ability to control database producers' data[6]. These effectively act as *de facto* laws that give copyright owners the ability to overcome the limitations of their government granted monopolies, undermining interoperation. They take a wide variety of forms. Passwords and internet protocol filtering allow the database owner to limit access to specific users and computers, and to selectively cut off access to researchers performing bulk calculations. Data can also be presented piecemeal, in response to a specific user query, thus limiting bulk downloads or incorporation into large-scale calculations. Databases can be stored in propriety formats, requiring users to view data through special software. Going a step further, the data can be encrypted, requiring the user to have a special code before it can be used. Effectively, propriety formats and encryption encumber the transfer of information to a medium where it can be manipulated and analyzed. Finally, watermarking adds overt or hidden digital fingerprints, slightly corrupting the data. It can prevent copying but it also adds background noise to large-scale calculations, potentially leading to errors. Examples of the application of these protections include the Incyte (Palo Alto, CA, USA) Proteome database (http://www.proteome.com/YPDhome.html) and the Cellzome (Heidelberg, Germany) database of interactions (http://yeast.cellzome.com/). These allow only users with passwords, sometimes filtered by an internet protocol address, to access pages that present data only in response to a limited query, preventing large-scale, global analysis. To date, watermarking of data does not seem to be a common mechanism in biological databases, although it can be found in other online resources, such as the British Library (London, UK).

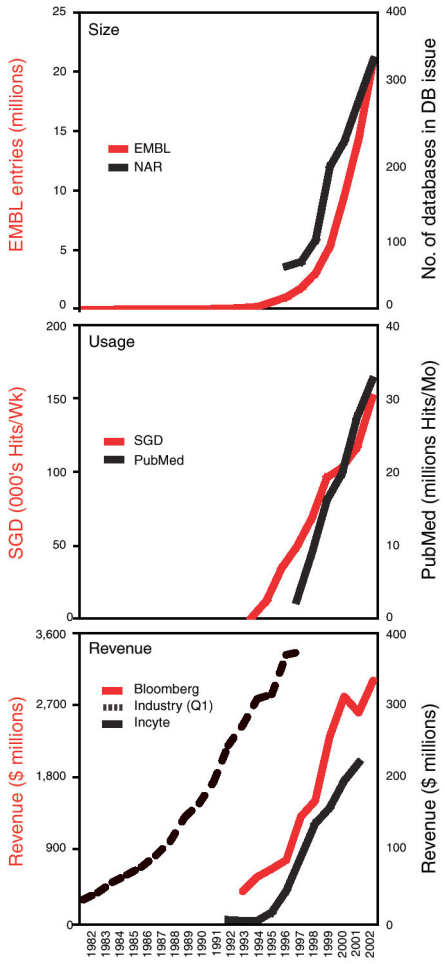

Figure 1 The historical and present state of the database sector. (**a**) Growth of biological databases over the past decade exemplified by the number of entries into the European Molecular Biology Laboratory (EMBL) database (http://www.ebi.ac.uk/embl/) and the number of databases described annually in Nucleic Acids Research (NAR)[10]. (**b**) Usage of these databases, taking as examples weekly hit data from the *Saccharomyces* Genome Database (SGD; http://www.stanford.edu/usage/sgd/) and monthly hit data from PubMed (http://www.ncbi.nlm.nih.gov) to show a corresponding exponential growth in the usage of databases. (**c**) Strong growth in revenue for the database sector. First quarter revenue for the industry from 1982 to 1997 (ref. 3). After 1997, we use two representative companies as examples: Bloomberg (a financial database company) and Incyte (a biological database company). Source: http://www.hoovers.com.

The US Digital Millennium Copyright Act (DMCA) strengthens these protections by making it illegal to surmount digital safeguards. Furthermore, the Anticounterfeiting Amendments (proposed in 2002 and expected to be reintroduced in the 108th Congress) would prohibit the extraction of digitally watermarked facts from their database. These bills both attempt to expand the scope of protection for databases in the US and provide for greater criminal and civil penalties, even for acts previously permitted—fair-use extraction of data for research—under copyright law. Moreover, the present excess and irregular application of these technologies serves only, from scientists' perspectives, to balkanize the database industry through over-fencing, creating a chaotic hodgepodge of individual fiefdoms.

Future ideas

The lack of cohesion between scientific databases, partially stemming from the diverse structure and organization of independently produced data sets, creates an impractical situation for integration. Even at the most basic level, databases tend to have incongruous structures: simple flat files, relational tables or object-oriented modules. There have been some moves among scientists toward creating systematic ontologies and standards for structuring databases (*e.g.*, refs. 7,8), but far more along these lines is needed.

Unfortunately, technological safeguards make this problem considerably worse. We feel the laws should discourage, rather than promote, these solutions to database protection. One solution that we support is a universal industry-wide standardized and compulsory license that would allow aca-

demic users the ability to access any data set at a reasonable price without having to negotiate different complex and limiting agreements for each database. These licenses can be designed with the normative methods of research in mind, including sharing of data and the ability to conduct bulk downloads. Compulsory licenses already exist in many other intellectual property spheres, notably the music industry (*e.g.,* Madonna did not have to ask permission from Don McLean, rather she had the option to employ a compulsory license to remake the song American Pie).

One of the advantages of this idea is that a standard framework for legal 'code' would obviously help promote standard computer code and interfaces. However, a potential problem is the possible devaluation of for-profit databases. Why would commercial organizations purchase databases at a significant cost when they could get them from academic sources that have accessed the data through the compulsory license scheme? To prevent such arbitrage, the law could require a time embargo on open access. When data first comes out and has its highest value, it would not be subject to the compulsory license (*e.g.*, ref. 9), which would only take effect after some fixed period.

In a direct effort to limit the use of technological safeguards (viewed as a necessary evil that, if databases owners were confident in the protections granted by copyright, would not be necessary), we propose that the law mandate that databases adhere to interoperability principles and limit technological protections as a prerequisite to attaining intellectual property protection. This is similar in spirit to the situation in patents where inventors are provided legal protection for an idea in exchange for publishing and completely revealing their methods.

The *status quo* shaped by international policy and commercial interests, although not limiting database growth, restricts science.

In conclusion, the *status quo* shaped by international policy and commercial interests, although not limiting database growth, restricts science; open access and universal interoperability is necessary for research. New legislation voicing the needs of scientific research is required. This legislation ought to promote research—through compulsory licensing and limiting technological safeguards—as well as promoting database creation through simply and uniformly protecting investment in databases.

1. *Feist Publications, Inc. v. Rural Telephone Service Company*, 499 U.S. 340 (1991).
2. *Pro CD, Inc. v. Zeidenberg*, 86 F.3d, 1447 (7th Cir. 1996).
3. Maurer, S.M., Hugenholtz, P.B. & Onsrud, H.J. *Science* **294**, 789–790 (2001).
4. Williams, M. *Information Market Indicators: Information Center/Library Market* (Market Indicators, Inc., Monticello, New Jersey; 1999).
5. Tenopir, C.B., Baker, G. & Robinson, W. *Library J.* **127**, 42–49 (2002).
6. Maurer, S.M. & Scotchmer, S. *Science* **284**, 1129–1130 (1999).
7. Ashburner, M. *et al. Nat. Genet* **25**, 25–29 (2000).
8. Lan, N., Montelione, G. T. & Gerstein, M. *Curr. Opin. Chem. Biol.* **7**, 44–54 (2003).
9. Patrinos, A. & Drell, D. *Nature* **417**, 589–590 (2002).
10. Database Issue. *Nucleic Acids Res.* 24–31 (January issue, 1996–2003).