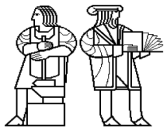

Massachusetts Institute of Technology
Engineering Systems Division

Working Paper Series

ESD-WP-2006-17

REUTILIZATION AND LEGAL PROTECTION OF NON-COPYRIGHTABLE DATABASE CONTENTS

Hongwei Zhu[1] and Stuart Madnick[2]

[1]**Massachusetts Institute of Technology**
mrzhu@mit.edu

[2]**Massachusetts Institute of Technology**
smadnick@mit.edu

August 2006

REUTILIZATION AND LEGAL PROTECTION OF NON-COPYRIGHTABLE DATABASE CONTENTS

Hongwei Zhu
College of Business & Public Administration
Old Dominion University
2147 Constant Hall
Norfolk, VA 23529
USA
hzhu@odu.edu

Stuart E. Madnick
Sloan School of Management
Massachusetts Institute of Technology
30 Wadsworth Street, E53-321
Cambridge, MA 02142
USA
smadnick@mit.edu

ABSTRACT
The availability of data on the web and the improvement of technologies have made it increasingly easy to reuse existing data to create new databases and provide value-added services. Meanwhile, initial database creators have been seeking legal protection for their data. After presenting a brief history of legislation related to legal protection for non-copyrightable database contents, we discuss challenging issues to be considered in formulating a database protection regulation. These issues can be addressed from the perspective of economics. Results from a preliminary economic analysis are presented. The findings indicate that depending on investment required to create the initial database and the level of differentiation between the initial database and the reuser database, the choice of a social welfare-enhancing regulation can allow for no reuse, free reuse, or fee-paying reuse.

KEY WORDS
database protection, data reuse, economic analysis

1. Introduction

The Web has become the largest data repository on the planet[1]. One of the main factors contributing to the success of the Web is its openness and ease of use: anyone can contribute data to, and consume data from, the Web. As Tim Berners-Lee, the inventor of the Web, put it in an interview by *Technology Review* in 2004, "the exciting thing is serendipitous reuse of data: one person puts data up there for one thing, and another person uses it another way". Such serendipitous data reuse is valuable and even vital to the development of the Web and the emerging Semantic Web. Through reuse, new knowledge can be created, innovation and value-added services become possible.

While we enjoy the availability of data and gain increasing capability of reusing data to create value, certain data reusers have been legally challenged. For example, *Bidder's Edge*, a web data reuser that created an integrated database by extracting data from more than 100 auction sites to facilitate easy search of auction items and prices, was sued by *eBay* in 1999. Meanwhile, there has been development on the legislative front to regulate data reuse activities. The European Union (EU) has introduced the Database Directive to restrict unauthorized data extraction and data reutilization. In the U.S., six bills have been introduced to provide legal protection for database contents - none of the six bills were passed into law. The significant uncertainty in the U.S. and the apparent divergence between the U.S. and the EU in database legislation have created serious challenges to the "serendipitous reuse of data".

Many computing professionals develop technologies (e.g., web wrapping, web services, and various Semantic Web technologies) to make data reuse much easier and more effective. It is important that computing professionals are aware of the legal implications when these technologies are applied for data reuse purposes. We will discuss these implications and the issues involved in providing legal protection for database contents.

2. eBay *v.* Bidder's Edge: Data Reusers Face Legal Challenges

With millions of items auctioned at hundreds of online auction sites, it can be time consuming to find the specific items of interest and keep track of their bidding prices. A number of auction data aggregators emerged to address the challenge. The aggregators employed computer agents to visit auction sites repeatedly and extract data automatically. Bidder's Edge was such an aggregator. It gathered bidding data of over five million items from more than 100 online auction sites, including eBay's. Bidder's Edge made it much easier to search and compare auction data across multiple sites. However, eBay was

[1] In the ensuing discussion, we will consider a website owner as a database creator.

concerned with the data extraction and reuse activities of Bidder's Edge. In late 1999, eBay sued Bidder's Edge and won a preliminary injunction in the following year based on a controversial interpretation of trespass law in the Internet context [2,11]. The case was settled later without a court decision.

There have been several other cases involving data reuse. A common characteristic in these cases is that the initial data compilers or the database creators (e.g., eBay) tend to be large and established firms, whereas the data reusers (e.g., Bidder's Edge) tend to be smaller firms using new technologies to extract and reuse data from the creator databases. Although there has been no definite answer to what types of data reuse are legal, certain data reusers stopped their activities in fear of the legal threats posed by the creators. Existing and emerging technology-enabled data reusers continue to face legal challenges. For example, data reusers that provide airfare comparison services have received warning letters from some online travel agencies (See "Cheap-Tickets Sites Try New Tactics" by A. Johnson, Wall Street Journal, October 26, 2004).

These cases have raised several questions regarding technology-enabled data reuse: Is it legal? Should it be regulated? If so, what are the issues and how should it be regulated? We will address these questions in the rest of the paper.

3. Feist *v.* Rural: Non-Creative Database Contents Are Not Copyrightable

The immediate reaction of many people is that most websites are copyrighted, thus extraction and reutilization of the data from these sites may violate copyright law. It turns out it is not the case. When it comes to data, copyright in the U.S. protects the original selection and arrangement of data, not the data itself or the effort in compiling the database. This principle was established in a landmark Supreme Court case between *Feist Publications* and *Rural Telephone Co* (499 US 340, 1991).

In compiling its phone book covering the service area of Rural, Feist copied about 8,000 records of Rural's White Pages. In the appeal case, the Supreme Court decided that Feist did not infringe Rural's copyright in that white pages lack the minimal originality to warrant copyright protection. This was because arranging entries alphabetically does not require any creativity.

Copyright law may evolve and play an important role in database protection in the future, but currently it does not restrict the reuse of the contents in many databases on the Web.

4. History of Database Legislation

Database creators have tried several ways to protect their non-copyrightable database contents. A commonly practiced protection is through access control, which often requires user subscription and authentication. But this does not prevent data extraction if the user provides identification to the aggregator (e.g., financial account aggregators [10].) Enforceable contracts to restrict the extraction and reuse of the data are difficult to establish on the Web unless cumbersome click-throughs are in place. This situation lets database creators feel that there exists a gap in the existing law which gives them no protection to their data and their investment in creating databases. As a result, they have started to seek means to protect their data through legislation.

The European Union (EU) first introduced the Database Directive in 1996 to provide legal protection for database contents and to safeguard the investment of database makers. Under its reciprocity provision, databases from countries that do not offer similar protection to databases created by EU nationals are not protected by the Directive within the EU. This created a situation where U.S. database creators felt they were neither adequately protected at home, nor abroad. In response, the database industry pushed the Congress to create a new law to provide similar protection to database contents. Since then, the U.S. has attempted six proposals, all of which already failed to pass into law. Figure 1 briefly summarizes these legislative proposals.

The *sui generis*[2] right approach taken by the EU creates a new type of right in database contents; unauthorized extraction and reutilization of the data is an infringement of this right. Lawful users are restricted not to "perform acts which conflict with normal exploitation of the database or unreasonably prejudice the legitimate interests of the maker of the database". Here "the legitimate interests" can be broadly interpreted and may not be limited to commercial interests.

HR 3531 of 1996 closely followed the EU approach with even more stringent restrictions on data reuse. Although the Database Directive has been adopted by the EU, HR 3531 failed in the U.S. One of the main concerns is the constitutionality of the scope and strength of the kind of protection in the EU Database Directive [3,9,12]. Other issues in the EU Database Directive include the ambiguity about the minimal level of investment required to qualify for protection [7,13], its lack of compulsory license provisions [3], the potential of providing perpetual protection under its provision of automatic right renewal after substantial database update, the ambiguity in what constitutes a "substantial" update, and several other issues which we discuss in the next section.

[2] In Latin, meaning "of its own kind", "unique".

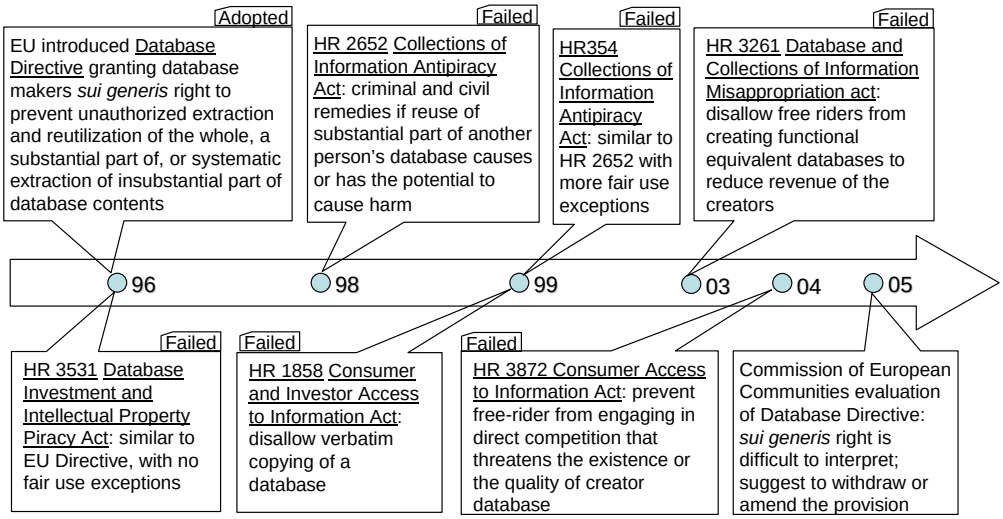

Figure 1. History of Database Protection Legislation

All subsequent U.S. proposals took a misappropriation approach where the commercial value of databases is explicitly considered. HR 2562 of 1998 and its successor HR 354 of 1999 penalize the commercial reutilization of a substantial part of a database if the reutilization causes harm in the primary or any intended market of the database creator. The protection afforded by these proposals can be expansive when "intended market" is interpreted broadly by the creator. At the other end of the spectrum, HR 1858 of 1999 only prevents someone from duplicating a database and selling the duplicate in competition.

HR 3261 of 2003 has provisions that lie in between the extremes of previous proposals. It makes a data reuser liable for "making available in commerce" a substantial part of another person's database if "(1) the database was generated, gathered, or maintained through a substantial expenditure of financial resources or time; (2) the unauthorized making available in commerce occurs in a time sensitive manner and inflicts injury on the database or a product or service offering access to multiple databases; and (3) the ability of other parties to free ride on the efforts of the plaintiff would so reduce the incentive to produce the product or service that its existence or quality would be substantially threatened". The term ''inflicts an injury'' means "serving as a functional equivalent in the same market as the database in a manner that causes the displacement, or the disruption of the sources, of sales, licenses, advertising, or other revenue".

The purpose of HR 3872 is to prevent misappropriation while ensuring adequate access to factual information. It disallows only the free-riding that endangers the existence or the quality of the creator database. Unlike in HR 3261, injury in the form of decreased revenue alone is not an offence.

On December 12, 2005, the Commission of European Communities [4] issued its first evaluation of the Database Directive. The evaluation shows that although the Directive helped harmonize copyright laws within the EU, the economic impact of the *sui generis* right on database production within the EU is unproven. In addition, the scope of the *sui generis* right has proved to the difficult to interpret and its related provisions have "caused considerable legal uncertainty, both at the EU and national level". Three policy options were offered in the evaluation report: repeal the whole Directive, withdraw the *sui generis* right, or amend the *sui generis* provisions. Similar recommendations appeared much earlier from research and education communities in Europe [5].

These legislative initiatives demonstrate the substantial difficulties in formulating a database protection law that strikes the right balance, which is a central issue in dealing with various kinds of intellectual property [1]. This issue is often manifested by several concerns, which we briefly discuss in the following section.

5. Concerns of Providing Legal Protection for Database Contents

Data monopoly. There are situations where data can only come from a sole source due to economy of scale in database creation or impossibility of duplicating the event that generates the data set. For example, no one else but eBay itself can generate the bidding data of items auctioned on eBay. A law that prevents others from using the factual data from a sole source in effect legalizes a data monopoly. Downstream value creating reutilizations of the data will be endangered by a legal monopoly. The European Court of Justice (ECJ) partially addressed this issue by trying to distinguish *data created* from *data obtained*, and by protecting only databases whose data is obtained [4].

Cost distortion. Both the EU database directive and the latest U.S. proposals require substantial expenditure in creating the database for it to be qualified for protection. Database creators thus may over invest at an inefficient level to qualify [14].

Update distortion and eternal protection. This is an issue in EU law, which allows for automatic renewal of *sui generis* right once the database is substantially updated. Such a provision can induce socially inefficient updates and make possible eternal right through frequent updates [8].

Constitutionality. Although the Congress in the U.S. is empowered by the Constitution to regulate interstate commerce under the Commerce Clause[3] and the misappropriation approach often gives a database law a commercial guise, the restrictions of the Intellectual Property Clause[4] often apply to any grant of exclusive rights in intangibles that diminishes access to public domain and imposes significant costs on consumers [6]. Most database contents are facts in the public domain; disallowing mere extraction for value creating activities runs afoul of the very purpose of the Intellectual Property Clause that is to "promote the Progress of Science and useful Arts". Excessive restrictions on data reuse may also violate the Constitution's First Amendment [5]. Since little extra value for the society as a whole is created by simply duplicating a database in its entirety, preventing verbatim copying of a database is clearly constitutional. Extracting all contents of a database is very much like duplicating the database. Unlike in copyright law where there is a reasonably clear idea-expression dichotomy (i.e., copyright protects the expression, not the idea conveyed by the expression), the constitutional line-drawing between extraction and duplication in data reuse is very difficult [6]. A constitutional database law needs to determine up to how much one is allowed to extract database contents.

International harmonization. Given the global reach of the Web and increasing international trade, it is desirable to have a harmonized data reuse policy across jurisdictions worldwide. The EU and the U.S. are diverging in their approaches to formulating data reuse policies. A World Intellectual Property Organization (WIPO) study [15] also reveals different opinions from other countries and regions.

[3] Constitution 1.8.3, "To regulate Commerce with foreign Nations, and among the several States, and with the Indian Tribes".

[4] Constitution 1.8.8, "To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries".

We believe the solution to these challenges hinges upon finding a reasonable balance between protection of incentives and promotion of value creation through data reuse. With this balance, value creation through data reuse is maximally allowed to the extent that the creators still have enough incentives to create the databases. Consensus can develop for international harmonization if we can determine the policy choices that maximize social welfare; a database policy so formulated should survive the scrutiny of constitutionality; other inefficiencies can be avoided or mitigated.

6. Achieving Balance in Database Legislation

We approach the challenge with an economic analysis, which considers the commercial value of databases. The details of the model can be found in [16]. In addition to the factors discussed below, the analysis in [16] also considers deficiencies in policy administration and discusses cost distortions where a database creator may over invest to gain maximum protection. Note that the policy choices suggested by the model are based on social welfare enhancement, not on fairness to any particular party (e.g., the creator, the reuser, or the consumer).

The analysis is based on competition of differentiated database products. It considers a database creator, which incurs a cost to create the initial database, and a data reuser, which extracts a certain amount of data from the creator database to create the reuser database. The reuser database can be differentiated from the creator database in terms of scope (e.g., extracting a fraction of the creator's data, combining it with data from other sources) and functionality (e.g., faster search algorithms). The reuser uses technology to allow it to easily extract and combine data from existing databases so that the cost of creating the reuser database can be negligible.

The competition from the reuser database can reduce the creator's revenue. When the reduction is such that the creator's revenue cannot offset its cost of creating the database, market fails and policy intervention is needed. This is the case where regulation for data reutilization is needed, either by creating a new law or by amending existing laws.

A regulation potentially can restrict certain stakeholders and benefit certain other stakeholders, but the society as a whole should better off with the regulation. That is, the choices of the regulation should enhance social welfare. Our analysis shows that such choices depend on the relationship among several factors. The most important two are the cost of creating the initial database and the level of differentiation between the creator database and the reuser database. Welfare-enhancing choices in relation to the two factors are depicted in Figure 2.

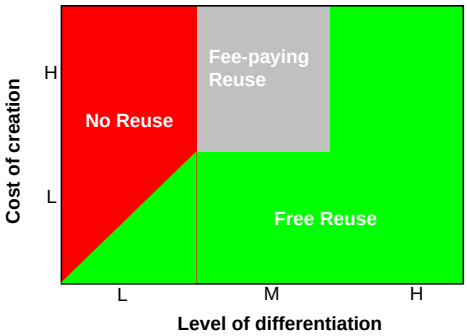

Figure 2. Welfare-enhancing Choices

When the level of differentiation is low, no reuse should be allowed. This is because such reuse adds little value to the society, at the same time the intense competition can drive the price so low that the creator cannot sustain even if the regulation requires the reuser to pay the creator a fee. Verbatim copying of an entire database is an extreme example of this scenario.

When the level the differentiation is moderate or high, there are two scenarios where it is welfare-enhancing to allow for free reuse: creation cost is low, or differentiation is high regardless of creation cost. With moderate differentiation, competition is not as intense as that in the case of low level of differentiation. The softened competition allows the creator to make enough revenue to offset its cost. When the level of differentiation is high, there will be little competition between the creator and the reuser. In other words, the data reutilization has little impact on the creator.

Although in both cases the regulation could require the reuser to pay the creator a fee, from social welfare point of view this is not desirable because there is always an inefficiency associated with money transfer, which is known as transaction cost. The fee can benefit the creator, but it does not create any extra value and the society as a whole incurs a transaction cost.

When the level of differentiation is moderate but the cost of creation is high, the reuser should pay a fee to the creator. This is the case where without a fee the reuse would cause market failure, but with a fee the creator can sustain. Since the creator may not be willing to license its data to the reuser, a compulsory licensing provision should be in place.

7. Conclusion

We have discussed legal issues related to technology-enabled data reuse, which makes many value-added services possible. But certain reuse may diminish the incentives of creating the initial databases, in which case intervention is needed to balance different interests. We presented the results of a preliminary study on how to balance the public interests of data reuse and the private of interests of profiting from creating the initial databases. The results show there is not a one-size-fits-all formula for data reuse regulation. Rather, depending on several factors, no reuse, free reuse, or fee-paying reuse should be allowed.

The implications of our preliminary results can be illustrated by revisiting the *eBay v. Bidder's Edge* case. According to our analysis, we need to at least examine the level of differentiation of the database developed by the reuser Bidder's Edge. In terms of searching of bidding data, the reuser database has a much broader coverage; thus, there is competition from the reuser database. In terms of functionality, eBay's database allows one to buy and sell items; the reuser database does not provide any actual auction service. Thus the two databases exhibit significant differentiation. Searching alone does not, in general, reduce eBay's revenue from its auction service. In addition, searching and actual auction are two different markets. If we subscribe to the spin-off theory [7], the eBay database will not meet the cost criterion. Therefore, free reuse by Bidder's Edge should be allowed.

As technologies for reusing data from various sources continue to emerge and improve, the need for understanding the legal implications of applying these technologies will become increasingly acute. A clear understanding of the implications will help legislators to formulate regulations with which serendipitous and innovative data reutilization can continue to provide value-added services without diminishing the incentives of compiling databases.

References

1. S. Bensen, L. Raskind, An Introduction to the Law and Economics of Intellectual Property, *Journal of Economic Perspectives*, **5**(1), 1991, 3-27.
2. D. L. Burk, The Trouble with Trespass, *Journal of Small & Emerging Business Law*, **4**(1), 2000, 27-56.
3. C. Colsten. Sui Generis Database Right: Ripe for Review? *The Journal of Information, Law and Technology*. **3**, 2001.
4. Commission of the European Communities (CEC), First Evaluation of Directive 96/9/EC on the Legal Protection of Databases. 12 December 2005, Brussels.
5. J. Grove, Wanted: Public Policies That Foster Creation of Knowledge. *Communications of the ACM*, **47**(5), 2004, 23-25.
6. P.J. Heald, The Extraction/Duplication Dichotomy: Constitutional Line Drawing in the Database Debate. *Ohio State Law Journal*. **62**(2), 2001, 933-944.
7. P.B. Hugenholtz, Program Schedules, Event Data and Telephone Subscriber Listings under the Database Directive: The "Spin-Off" Doctrine in the Netherlands and elsewhere in Europe. 11th Annual Conference on International Law & Policy, New York, 2003.

8. C. Koboldt, The EU-Directive on the legal protection of databases and the incentives to update: An economic analysis. *International Review of Law and Economics* **17**(1), 1997, 127-138.
9. J. Lipton, Private Rights and Public Policies: Reconceptualizing Property in Databases. *Berkeley Technology Law Journal* **18**(3), 2003, 773-852.
10. S.E. Madnick, M. D. Siegel, Seize the Opportunity: Exploiting Web Aggregation. *MISQ Executive* **1**(1), 2002, 35-46.
11. M. A. O'Rourke, Shaping Competition on the Internet: Who Owns Product and Pricing Information? Vanderbilt Law Review, **53**(6), 2000, 1965-2006.
12. J.H. Reichman, P. Samuelson, Intellectual Property Rights in Data? *Vanderbilt Law Review*, **50**(1), 1997, 52-166.
13. H.G. Ruse, Electronic Agents and the Legal Protection of Non-creative Databases, *International Journal of Law and Information Technology*, **3**(3), 2001, 295-326.
14. P. Samuelson, Legal Protection of Database Contents. *Communications of the ACM*, **39**(12), 1996, 17-23.
15. H. Tabuchi, International Protection of Non-Original Databases: Studies on the Economic Impact of the Intellectual Property Protection of Non-Original Databases. CODATA 2002, Montreal, Canada, 2002.
16. H. Zhu, S.E. Madnick, M.D. Siegel, Policy for the Protection and Reuse of Non-Copyrightable Database Contents. MIT Sloan School Working Paper #4751-05, 2005.