One Size does not Fit All:
Legal Protection for Non-Copyrightable Data

Hongwei Zhu
Stuart E. Madnick

Working Paper CISL# 2007-04

July 2007

Composite Information Systems Laboratory (CISL)
Sloan School of Management, Room E53-320
Massachusetts Institute of Technology
Cambridge, MA 02142

One Size does not Fit All: Legal Protection for Non-Copyrightable Data

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

Introduction

The Web has become the largest data repository on the planet[1]. An important factor contributing to its success is its openness and ease of use: anyone can contribute data to, and consume data from, the Web. As Tim Berners-Lee, inventor of the Web, said[2], "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 extremely valuable. Through reuse, new knowledge can be created, innovation and value-added services become possible.

However, there have been efforts to regulate and legally challenge data reuse activities. The European Union (EU) has adopted the Database Directive to restrict unauthorized data extraction and reuse. In the U.S., Congress has considered six bills, all of which failed to pass into law. These legislative activities are summarized in Figure 1; more details are furnished later. The significant uncertainty and the international differences in database legislation have created serious challenges to the "serendipitous reuse of data". The dual purposes of this paper, both related to the theme "one size does not fit all", are to: (1) summarize the range of legislation in current use and proposed, and (2) present an economic model for interpreting or recommending policy choices that depend on factors such as cost of database creation and level of database differentiation.

[1] In the ensuing discussion, we will consider a website owner as a database creator.
[2] An interview by *Technology Review*, October, 2004, p44.

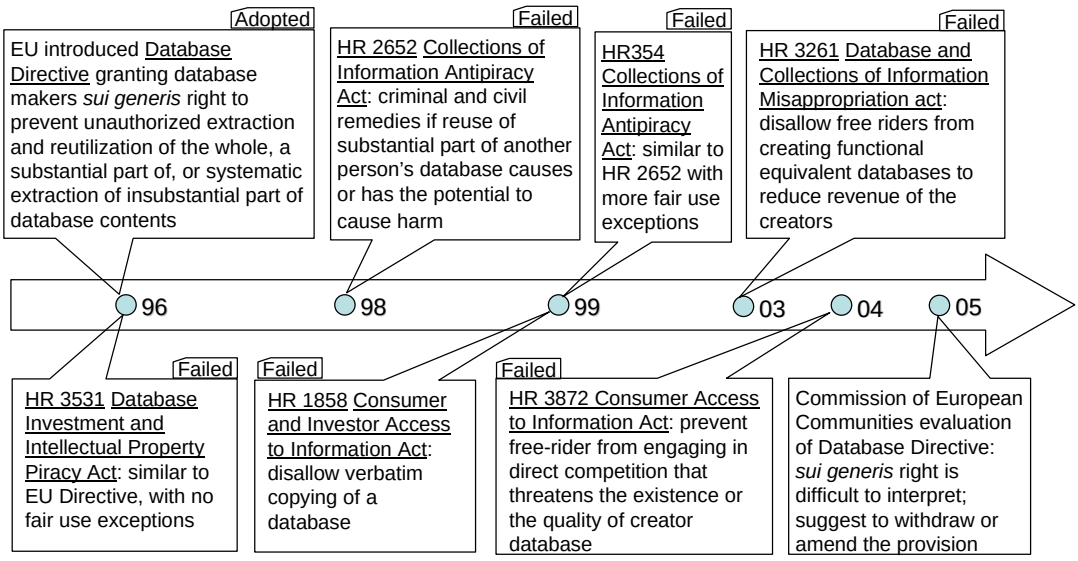

Figure 1. History of Database Protection Legislation

As computing professionals continue to develop technologies (e.g., data extract, web mashups, web services, and various Semantic Web technologies) to make data reuse much easier, it is important for us to understand the legal implications when applying these technologies for data reuse purposes.

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

Let us start with an example. 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 on multiple auction sites. A number of auction data aggregators, such as Bidder's Edge, emerged to address the challenge by employing computer agents to visit auction sites repeatedly and extract data systematically. Bidder's Edge made search and comparison of auction data across multiple sites much easier by gathering bidding data of over five million items from more than 100 online auction sites, including eBay. However, 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 [9]. The case was settled later without a court decision; Bidder's Edge ceased operation and the company no longer exists.

There have been several other cases involving data reuse in the U.S. A common characteristic in these cases is that the data reusers (e.g., Bidder's Edge) tend to be smaller firms using new technologies to extract and reuse data from one or more creator databases. In many cases, the 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[3].

Data reusers in Europe have also faced legal challenges. For example, William Hill, an online betting company in the U.K., created a database by combining its own data (e.g., betting odds) with horse racing event data published by British Horseracing Board (BHB), which is the governing authority for organizing horse races in the U.K. William Hill displayed the contents of the database on its website to facilitate its betting business, but was sued by BHB for its systematic reuse of BHB's data.

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.

Feist *v.* Rural: Non-Creative Database Contents Are Not Copyrightable in the U.S.

Many people think that the factual data on websites is copyrighted, thus extraction and reuse of the data from websites is well-defined and controlled by copyright law. It turns out that is not the case.

[3] See "Cheap-Tickets Sites Try New Tactics" by A. Johnson, Wall Street Journal, October 26, 2004.

When it comes to data, copyright in the U.S.[4] protects the original selection and arrangement of data, but 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*.[5] In compiling its phone book covering the service area of Rural, Feist reused 1,309 of the approximately 7,700 listings in Rural's White Pages. In the appeal case, the Supreme Court decided that Feist did not infringe Rural's copyright in that Rural's white pages lack the requisite originality to warrant copyright protection. Originality requires a work to be "independently created by the author" and it must possess "at least some minimal degree of creativity". Arranging entries alphabetically does not have the required degree of creativity.

The Court confirmed that "copyright rewards originality", originality requires "some minimal degree of creativity", and "Originality is a constitutional requirement." It also rejected the so-called "sweat of the brow" doctrine that considers copyright as a "reward for the hard work that went into compiling facts." The implication of this landmark decision is that in the U.S. copyright currently does not restrict the reuse of the factual contents in most publicly accessible databases on the Web[6].

The Court decision, together with the exponential growth of digital information and the increasing technological capability of reusing information, have induced a series of legislative activities to provide legal protection for database contents.

Internationally Copyright Provides Differing Degrees of Protection to Databases

Copyright law differs internationally in terms of how much protection it extends to factual databases. In the U.S., copyright protects the creative selection and arrangement of data, not the

[4] International differences are discussed later.
[5] U.S. Supreme Court, 499 US 340, 1991.
[6] Note that Web content, such as news articles, music, video, and such, are not data and are protected by copyright law. The focus of this article is on data – such as, in the previous example, the list of items for sale on eBay and their auction prices.

data itself. In other words, the creative choice of what to be included in a database and the creative design of the database schema are protected by copyright in the U.S., but not the factual data records.

Although the U.S. has rejected the "sweat of the brow" doctrine, Australia embraces the doctrine for its copyright law as evidenced by the appeal case *Desktop Marketing Systems Pty Ltd v. Telstra Corporation Limited*[7]. Desktop used all the entries in Telstra's white pages and yellow pages to make CD-ROMs with several additional search features. The Full Court ruled that originality "does not require novelty, inventiveness or creativity", and a work is original "if the compiler has undertaking substantial labour or incurred substantial expense in collecting the information recorded in the compilation." The High Court of Australia confirmed the judgment in 2003 and maintained that Desktop infringed Teslstra's copyright.

The different creativity requirements of the U.S. and Australia represent two extremes. The Canadian law is somewhere in between the extremes. In the judgment of a Canadian case[8], the Court decided that originality "need not be creative, in the sense of being novel or unique." A work is original if it is "more than a mere copy of another work" and requires "an exercise of skill and judgment" that "must not be so trivial that it would be characterized as a purely mechanical exercise."

Despite these differences in the criteria for testing originality, copyright law is quite uniform internationally that one cannot claim copyright protection for individual entries of facts stored in a database.

[7] Full Federal Court of Australia, 2002.
[8] Supreme Court of Canada, CCH Canadian Ltd. V. Law Society of Upper Canada, 2004.

History of Database Legislation

Database creators have tried several ways to protect their non-copyrightable contents[9]. A commonly practiced method 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., a user provides login credentials to a financial account aggregator for it to gather information from disparate accounts on the user's behalf [8].) Enforceable contracts to restrict the extraction and reuse of the data are difficult to establish on the Web unless cumbersome "click-through" agreements are in place. As a result, some database creators feel existing law does not give them sufficient protection to their data and their investment in creating databases. Consequently, they have sought means to protect their data through new legislation. See Figure 1 earlier for a summary of legislative activities.

The EU first introduced the Database Directive in 1996 to provide two kinds of protection for a database: copyright for the selection or arrangement of database contents, and *sui generis*[10] right for the contents in the database. The *sui generis* right is a new type of right to prevent unauthorized extraction and/or reutilization of the whole, a substantial part, or systematic extraction and/or reutilization of an insubstantial part, of contents of a database that is created with substantial expenditure. 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.

[9] Due to limitations on length, we will not discuss all the technical methods that have been used, such as blocking requests from IP addresses that appear to be extracting large quantities of data, etc. In general, for each technical approach to prevent data extraction, there is a possible technical counter-measure to overcome it.
[10] In Latin, meaning "of its own kind", "unique".

The Directive has been criticized for its ambiguity about the minimal level of investment required to qualify for protection [5], its lack of compulsory license provisions [1], the potential of providing perpetual protection under its provision of automatic right renewal after a substantial database update, and the ambiguity in what constitutes a "substantial" update.

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. In response, the U.S. database industry pushed the Congress to provide similar protection to database contents. Since then, the Congress has considered six proposals, all of which have failed to pass into law.

HR 3531 of 1996 closely followed the EU Database Directive approach with even more stringent restrictions on data reuse. One of the main concerns is the constitutionality of the scope and strength of the kind of protection for database contents [1,7].

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.

Following the reasoning in the NBA v. Motorola case[11], 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

[11] 105 F.3d 841 (2nd Circuit, 1997). Motorola transcribed NBA playoff scores from broadcast and sent them to its pager subscribers. The misappropriation claim by NBA was dismissed.

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 [2] 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".

These world-wide legislative initiatives demonstrate the substantial difficulties in formulating a database protection law that balances creator incentives and the values added by data reuses. Some of the challenges are briefly discussed below.

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 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 which would endanger any downstream value-creating reutilizations of the data. 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 by collecting existing independent materials[12].

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 [10]; see [12] for an economic model that explains such cost distortion.

Update distortion and eternal protection. This is an issue in EU law, which allows for automatic renewal of *sui generis* right when the database has been substantially updated. Such a provision can induce socially inefficient updates solely to attain eternal rights [6].

Constitutionality. Although the U.S. Congress is empowered by the Constitution to regulate interstate commerce under the Commerce Clause[13] and the misappropriation approach often gives a database law a commercial guise, this must be balanced against the Intellectual

[12] European Court of Justice, Grand Chamber, The British Horseracing Board Ltd and Others v. William Hill Organization Ltd., 2004. A database creator with data that is *created*, e.g., BHB, which created the fixture list, would be a natural monopoly if legal protection was granted. Data that is *obtained* presumably could be obtained by anyone willing to make the effort.

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

Property Clause[14] which restricts the grant of exclusive rights in intangibles that diminishes access to public domain and imposes significant costs on consumers [4]. Certain database contents are factual data in the public domain; disallowing mere extraction of such data for value-creating activities runs afoul of the very purpose of the Intellectual Property Clause to "promote the Progress of Science and useful Arts". Excessive restrictions on reuse of factual data (a form of speech or press) may also violate the Constitution's First Amendment [3], which protects the freedom of speech and press. 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. A constitutional database law needs to determine how much one is allowed to extract database contents. The constitutional line-drawing between extraction and duplication in data reuse is very difficult [4].

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. We have discussed some of the differences in the U.S., the EU, Australia, and Canada. A World Intellectual Property Organization (WIPO) study [11] also reveals different opinions from other countries and regions.

A key element to solving these challenges hinges upon finding the right factors for 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 effectively

[14] 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".

balance these factors; a database policy so formulated should survive the scrutiny of constitutionality and other inefficiencies can be avoided or mitigated.

Achieving Balance in Database Legislation

We approach the challenge with an economic model [12] that considers the commercial value of databases. Based on differentiated competition theory, the model 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., different kind of search algorithm). 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, the market fails[15]. From an economic point view, regulation for data reutilization is needed to prevent or correct market failure.

A regulation potentially can restrict certain stakeholders and benefit certain other stakeholders, but the society as a whole should better off with the regulation. Our analysis shows that such choices depend on the relationship among several factors. The most important two are: (1) the cost of creating the initial database and (2) the level of differentiation between the creator database and the reuser database. The choices[16] in relation to these two factors are depicted in

[15] Market failure is an economic term for the situation where goods or services cannot be provided to consumers (e.g., it is not profitable for creator to produce the database.) Policy intervention can sometimes restore a failed market.

[16] There are actually more than three regions in our paper [12], we have simplified the situation slightly to shorten this paper.

Figure 2, which, as we mentioned earlier "one size does not fit all," illustrates that the policy choices are not just binary.

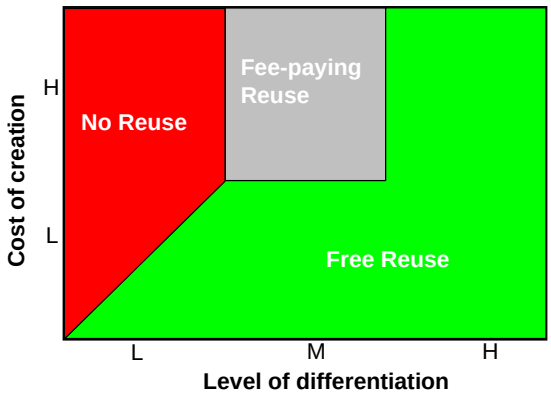

Figure 2. Policy Choices Suggested by the Economic Model

No reuse region. When the level of differentiation is low, not allowing reuse is a reasonable policy choice since such reuse adds little value, and, at the same time, the intense competition can drive the price so low that the creator cannot have enough revenue to offset the cost. Verbatim copying of an entire database is a typical example of this scenario.

Free reuse region. When the level of differentiation is moderate or high, there are two scenarios where free reuse should be allowed: 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 differentiation. The softened competition allows the creator to make enough revenue to offset its cost. With high differentiation, 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 reuser could be required to pay the creator a fee, this is not needed to prevent market failure and 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.

Fee-paying reuse region. 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.

Some Examples Illustrating the Application of these Principles

The economic model provides a useful framework for facilitating the ongoing debate of database legislation, analyzing data reuse cases, and interpreting court decisions. We will illustrate the applications of the model by revisiting the two cases mentioned earlier.

eBay v. Bidder's Edge. 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. eBay can still compete in the search space, but according to the model eBay should not be given the right to prevent innovative firms such as Bidder's Edge from offering search function before eBay acquires the necessary technical and business skills. Furthermore, if we subscribe to the spin-off theory [5], the eBay database will not meet the cost criterion. Therefore, free reuse by Bidder's Edge should be allowed.

BHB v. William Hill. The ECJ determined that although William Hill did systematically extract and reuse an insubstantial part of BHB's database, the cumulative effect has no possibility for William Hill to "reconstitute and make available to the public the whole and

substantial part of the contents of the BHB database" and therefore "seriously prejudice the investment" in the creation of the database. The criterion of "reconstitution" effect can be explained using the economic model as the reuser database having little differentiation. The ECJ also stressed that the injury needs to be serious, which can be understood from the market failure perspective in the model.

BHB spends £4 million annually to maintain the database. The ECJ judgment provides a guideline for determining if this cost is protected by the Database Directive. After making the distinction between creating and obtaining data, the ECJ determined that the investment protected by the *sui generis* right "does not cover the resources used for creating the materials which make up the contents of a database." To create the racing list, BHB had to verify information of participants, e.g., a horse's age and pedigree, and such information was *obtained* by BHB. The ECJ further ruled that "The resources used for verification during the stage of creation of materials" are not part of protected investment. These cost accounting rules used by the ECJ constitute a particular standard of determining the cost factor in the model.

Conclusion

Although the legislative efforts may seem to have stalled in the U.S. during the past two years, the issues related to technology-enabled data reuse have not been resolved. We discussed these issues and presented the preliminary results of an economic analysis on how to balance the benefits of data reuse to society and the interests of profiting from creating the initial databases[17]. 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 are welfare-enhancing choices.

[17] There are many other factors, such as the political, legal, and enforcement processes in different jurisdictions, that are beyond the scope of this paper. The intention of this paper is to establish some basic principles that could facilitate these other processes.

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. We are continuing to develop further understanding of the issues related to applying data reuse technologies. We anticipate the research to bring us closer to finding the right balance with which serendipitous and innovative data reutilization can be maximally allowed to provide value-added services without diminishing the incentives of compiling databases and making them available on the web.

References

1. Colsten, C. Sui Generis Database Right: Ripe for Review? *The Journal of Information, Law and Technology* 3 (2001).
2. Commission of the European Communities (CEC). First Evaluation of Directive 96/9/EC on the Legal Protection of Databases. December 12, 2005, Brussels.
3. Grove, J. Wanted: Public Policies That Foster Creation of Knowledge. *Communications of the ACM* 47, 5 (2004), 23-25.
4. Heald, P.J. The Extraction/Duplication Dichotomy: Constitutional Line Drawing in the Database Debate. *Ohio State Law Journal* 62, 2 (2001) 933-944.
5. Hugenholtz, P.B. 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 (2003), New York.
6. Koboldt, C. 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.
7. Lipton, J. Private Rights and Public Policies: Reconceptualizing Property in Databases. *Berkeley Technology Law Journal* 18, 3 (2003) 773-852.
8. Madnick, S.E., Siegel, M. D. Seize the Opportunity: Exploiting Web Aggregation. *MISQ Executive* 1, 1 (2002) 35-46.
9. O'Rourke, M.A. Is Virtual Trespass an Apt Analogy? *Communications of the ACM*, 44, 2 (2001), 98-103.
10. Samuelson, P. Legal Protection of Database Contents. *Communications of the ACM* 39, 12 (1996), 17-23.
11. Tabuchi, H. International Protection of Non-Original Databases: Studies on the Economic Impact of the Intellectual Property Protection of Non-Original Databases. CODATA (2002), Montreal, Canada. http://www.codata.org/codata02/03invited/Tabuchi/Tabuchi_CODATA_ejournal.pdf.
12. Zhu, H., Madnick, S.E., Siegel, M.D. Policy for the Protection and Reuse of Non-Copyrightable Database Contents. MIT Sloan School Working Paper (2005) #4751-05. Available at SSRN: http://ssrn.com/abstract=876960.