Attribution Principles for Data Integration: Technology and Policy Perspectives Part 2: Focus on Policy

Thomas Lee

CISL WP#02-04
February 2002

Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02142

Attribution Principles for Data Integration: Policy Perspectives

February 2002

Thomas Lee

Department of Operations and Information Management
University of Pennsylvania, The Wharton School

This page left intentionally blank.

Attribution Principles for Data Integration: Policy Perspectives

by
Thomas Y. Lee

Abstract
This paper is excerpted from a thesis submitted to the Engineering Systems Division in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the Massachusetts Institute of Technology in January of 2002. This paper addresses problems of attribution that arise from the data integration that is exemplified by data re-use and re-distribution on the Web. In an earlier document, we began with a simple definition of attribution, asking *what* data are we interested in and *where* does it come from? However, because the issue is more complex than simply *what* and *where*, we expand the scope of our analysis in this excerpt. From the perspective of intellectual property policies, we adopt a broader view of the attribution problem space. A policy analysis that surveys the status quo policy landscape and stakeholder interests is followed by specific policy recommendations. Informed by our technology perspective, we offer two new arguments to support misappropriation as a policy approach to the attribution problem space.

The policy perspective encompasses not only *what* and *where* but also integration architectures and the relationships between data providers and users. Information technologies separate the processes and products of data gathering from data selection and presentation. Where the latter is addressed by copyright, the former is not addressed at all. Based upon two traditional, legal-economic frameworks, the asymmetric Prisoner's Dilemma and Entitlement Theory, we argue for a policy of misappropriation to support integration and attribution for data.

Thesis Supervisor: Stuart E. Madnick
Title: John Norris Maguire Professor of Information Technology
MIT Sloan School of Management

1 Introduction

In the legend of Theseus, the hero of Athens entered the Labyrinth of Daedalus on the Isle of Crete to face the Minotaur. Critical to both his successful hunt and victorious return was the simple ball of thread that Theseus used to trace his path. (Bulfinch 2001; Lindemans 2000) As the wealth of content available via electronic networks continues to grow, the Internet has become a maze to rival Daedalus' Labyrinth.

Today, the World Wide Web is a popular way to access the Internet. One group of tools to help people navigate the labyrinth of on-line content are integration services that allow a user to pose rich queries across multiple sites and aggregation services which effectively roll several different sources behind a single point of entry (like Web portals). Consider for example, the case of planning a vacation. The Web may be like having the library on your desktop, but in at least one way, the virtual is no better than the physical. You still must go to the travel section (in the library or on some Web portal like Yahoo!™) and search the different travel guides.

Suppose that you are planning a trip to Japan. There are dozens of on-line resources, many accessible over the Web, ranging from guides for budget conscious travelers (Lonely Planet, Hostelling International) to more traditional guides (Frommer's Travel Guides) to application specific resources (Hotelguide.com, roomz.com). Note that these are resources for researching your trip. We are not discussing transactions such as making reservations or purchasing event tickets.

Rather than laboriously surfing through multiple guides, suppose that you had access to a Travel Resource Integrator (TRI). You might then want to ask:

> Q1 What places in Tokyo, Japan may a person traveling alone find a single bed for less than 25,000¥?

The TRI might provide you with the following table:

name	price
Asakusa View	18000
Ginza Dai-Ichi	15000
Dai-Ichi	10000
Grand Palace Hotel	10000
Asakusa Prince	10000
Hotel Sofitel	17000
Tokyo Yoyogi	3000
Tokyo International	3100
Sky Court Koiwa	4500
Sky Court Asakusa	5000

Table 1.1 Results for Q1

While demonstrating the convenience of such a tool, this example also serves to illustrate at least one specific problem with data integration tools like the TRI that applies not only to users but to providers of on-line resources such as those accessible over the Web. Specifically,

> Where does this information come from?

You as a user might like to know where the information comes from for reasons such as quality or search. Some questions related to quality that you might wonder include:

- Do you trust the source of this hotel list?
- Does this hotel list draw upon established, reputable resources such as Frommer's or Baedeker's, or is the list compiled from the memories of people who traveled to Tokyo twenty years ago?
- Is the information in the list current? Hotel prices often fluctuate significantly depending upon the time of year you wish to travel. Are all of the listed establishments still in business?

Even if you assumed the veracity of the content, once you had a list, you might want to read more about a specific hotel. To read additional information, you would want to look in the guide where you originally learned about the hotel in question. For example, you would want to know that the listing for the *Asakusa View* came from the Frommer's. Additional information that might be answered from the sources include:

- Are any on this list single beds (e.g. youth hostels) rather than single rooms?
- Which of these lodging options, if any, are located by interesting tourist attractions?
- How can I make a reservation at one of these listings? Is there a phone number to call?

Information providers also have an interest in knowing where information comes from and how data flows. Who should receive acknowledgement for preparing the data in your query result? Who should be paid for this data? If the information is older than the copyright term limit, is the content transferred to the public domain (and therefore free). However, how would individual users know which data fit that category? A single query, moreover, may use information from more than one place. How are rights and remuneration rationed between different contributors? The problem, for both users and the market as a whole, made difficult by the migration from physical to electronic, is only exacerbated by the Web, which makes it easy for people to link and frame or copy content from other sources.

In summary, then, we have suggested three general reasons why attribution is important: data quality, search, and intellectual property.

The question of attribution and its implications is not merely speculative. mySimon Inc. is a comparison shopping service that aggregates data from a number of on-line catalogs in a single data warehouse to facilitate user search. In 1999, mySimon brought suit against

Priceman, another comparison shopping service, charging, among other claims, that "Priceman did not sufficiently attribute its meta-search results to mySimon (Kaplan 1999)."

eBay, Inc. hosts an on-line auction house that allows users to play the parts of both buyer and seller. Sellers post items for auction in a database of products that buyers may browse or search and bid for. Bidder's Edge (BE), a comparison service not unlike mySimon or Priceman, warehoused the contents of several auction houses including eBay, Amazon, and Yahoo. eBay won a preliminary injunction against BE's practice in a lawsuit that included the complaint that "caching can lead to outdated information ... potentially harming eBay's reputation (Krebs 2000)."

While these two cases highlight the relevance of attribution-related issues, they also highlight a third point, the legal distinction between individual users and third party services. Suppose that eBay and mySimon were on-line travel resources. An individual user, like a physical shopper, could certainly have behaved like an integrator by visiting different stores and comparing prices without inducing any lawsuits. What if you asked a friend to shop for you, however? What if you paid a personal assistant to shop on your behalf? What about a commercial service? Finally, to what degree can the integration service "anticipate" your requests and search in advance? Ultimately, how far removed from an individual user can an integration service stray while still claiming to "stand in the shoes" of that user?

Details of these cases and others will be discussed further below. However, even this brief introduction serves to illustrate the tension generated by integration: Users benefit from integration, but integration can reduce a database producer's incentives to the point that there are no databases to integrate. As Senator DeWine explained, the threat is that "investment in databases will diminish over time.... Ultimately, the reliability of information available to consumers over the Internet would be undermined (MacMillan 2000)."

1.1 Technology and policy, an integrated approach

The thesis from which this paper is excerpted is about technologies and policies for balancing the tension between database integration and database production. Data integration is a challenging problem with issues that range from the technical (e.g. semantic and syntactic heterogeneity between sources (Goh 1997; Wiederhold 1992) to policy (e.g. standards for data organization and presentation (e.g. EDI, ASN.1, XML). In an earlier paper (Part 1 excerpted from the same thesis from which this paper is taken), we considered a formal model for a technology-based approach to documenting data sources. This paper, by contrast, identifies a set of intellectual property-related challenges to integration that stem from the problem of attribution (i.e. knowing where data comes from). Policy measures to both limit and support integration based upon where information comes from are considered. We consider both traditional and novel measures that judges and legislators have invoked to craft the current policy framework surrounding data integration technologies.

Before delving into the technology or the policy, the thesis describes the attribution-related problem space that stems from data integration. In the remainder of this Chapter, we sketch a broad outline of the problem space and operationally define attribution as a list of desiderata to address the problem space. Then, in Section 2, we provide a very brief overview of a number of the diverse, perhaps seemingly unrelated research streams that address this topic.

Section 3 is a policy analysis. We survey the current policy landscape by revisiting the initial desiderata from the broader, policy perspective. Then, we review the status quo legal framework addressing those issues, identify the stakeholders, and catalog their respective interests. Section 4 is a policy formulation exercise. We begin by clarifying the policy objectives and then redefine the problem in terms of technical database systems principles that are often overlooked in conventional policy exercises. We offer two theoretical frameworks, the Prisoner's Dilemma and Entitlement Theory, that are useful for evaluation and applicable to our problem redefinition. We present a specific proposal, a Federal misappropriations statute for data reuse and reintegration and evaluate that proposal in light of the frameworks.

Chapter 5 concludes this paper by comparing our policy formulation to the stakeholder interests in Chapter 4. As a part of the evaluation, we discuss both limitations of and proposed extensions to this research.

1.2 Scope

This paper, and the thesis from which it is drawn, is about technology and policy for data integration and attribution in the commercial market for use and reuse of data. However, not all types of data are treated in this analysis. Building from Tyson and Sherry (1997), we provide a brief taxonomy of different kinds of data to prescribe the scope of this research. The taxonomy can be thought of as defining a multi-dimensional space where each dimension describes the range of one type or category. Rarely is data, or its use, of a single, distinct type. Instead, a specific type or a specific use of data will often exhibit characteristics of multiple categories.

The first dimension of data that we consider is the initial purpose for which the data is gathered. Data collection might be driven by government mandate or by private interests. For example, a large body of financial performance figures is gathered in accordance with U.S. Federal reporting requirements. Telephone companies are required to assemble White page directories (Feist v. Rural 1991). Private organizations and associations collect other data including sports statistics (the National Basketball Association), academic ratings (U.S. News and World Report), and consumer buying habits (the New York Times Bestseller Lists). Individual collections of data range between the two extremes of government data and private interests.

A second dimension is the time sensitivity of the data distribution. Information often exhibits a "U" shaped value curve where value diminishes over time but eventually regains value in an archival context. Stock quotes are often cited as an example for which the timeliness of the

data strongly differentiates users (e.g. real-time for a fee vs. delayed for free). Real-estate listings, event listings, and travel guides represent other data that fall along the continuum of time sensitivity. In this dimension, data varies from being extremely time sensitive to being invariant.

Third, data may vary with respect to its replicability. Ignoring the question of whether it would be economically efficient to do so, is it possible for a second-comer to recreate the data set without resorting to any reuse of existing data? By its very nature, experimental scientific data is supposed to be replicable. However some data can neither be recreated nor gathered anyplace other than from its initial source. The current trading price of a stock on the New York Stock Exchange during trading hours is one such example. We therefore think of sole source data as not being replicable. The polar opposite is a data set that anyone can recreate.

We depict these three dimensions and their inter-relationships in Figure 1.1. We use the spheres (and their respective shadows) to illustrate how different types of data fit within the space. We might think of a 'Hotel price', for example, as being extremely *time sensitive.* Prices might change daily in response to changing demand. Moreover, prices from a single hotel come only from that hotel and so are considered *sole-source.* Barring false advertising claims, the government may have little interest in how a hotel chooses to advertise its prices. We do not think of government mandated publication of hotel price lists. The *purpose* for gathering or posting prices is therefore considered private. Next, we consider a U.S. Department of State Travel Advisory. Such warnings are issued by the government and may be based upon top-secret, national security related information. We may therefore think of Travel Advisories as *highly time-sensitive, sole source, government* data. In stark contrast, we consider a listing of publicly accessible tourist sites. Monuments and parks are unlikely to change over time and can be gathered and published by anyone. While the government may maintain such lists, there is no mandate enjoining or requiring competing private collections.

We might also think of a fourth dimension, that of individually identifiable information. Data that can be traced back to a specific individual raises the specter of privacy concerns. Because of the difficulty in illustrating four dimensions, we only show the interactions between three. In this thesis, we explicitly exclude consideration of data that falls into the spaces encompassed by government-sponsored, sole-source (non-reproducible), and individually identifiable data. Some of our analysis may apply more broadly. For example, attribution technology could apply to data gathered by government mandate. However, each of these categories also raises additional considerations, such as the policy management of individual privacy rights or the anti-trust provisions that stem from truly sole-source providers, that are considered outside the scope of this thesis.

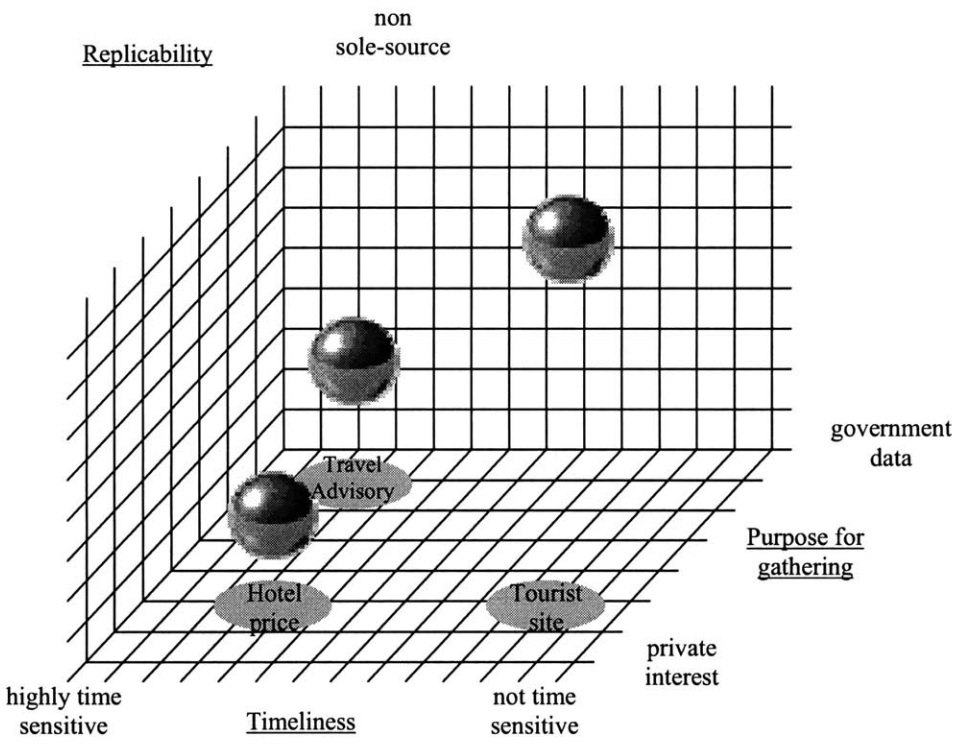

Figure 1.1 Three of the four dimensions of data

1.3 Integration challenges: the attribution problem space

We began this Chapter with a simple example to provide users with an intuition for what the term attribution means and to motivate the need for addressing attribution-related challenges to data integration. At that time, we informally defined attribution as some association between search results and the sources used to answer a particular question. Our goal now is to refine that intuition in two ways. First, we want to provide a broad outline of the problem space as a framework for tying together the technology analysis in Part 1 and the policy analysis in Part 2. Second, we will operationally define attribution as a list of desiderata for different attempts to address the space.

We begin by recalling some of the questions that any user of a data integration service might ask. We then provide a more systematic description of integration and ask what concerns a data provider might have about integration services. Finally, we assemble user and provider concerns into a general framework that defines the attribution-related integration problem space. From this characterization, we provide the list of desiderata.

1.3.1 User interests

Hearkening back to our initial motivating example, recall that we surmised that users of data integration services might be interested in general issues. First, they might like to know a bit more about the quality of the integrated information, and second, they might like to know where they could go to find additional corroborating or related information. More generally, we can characterize these two interests as questions about "*where* specific pieces of information (*what*) come from," and "*when* the information was gathered." By asking, "*what* information comes from *where*," and "*when* did we get that information," we begin to build the attribution problem space.

What addresses the issue of specificity. The answer to a single query may come in several parts. When asking about hotels in Tokyo, we might have consulted several different guidebooks. Because no single guide is necessarily exhaustive, different answers might have come from different guidebooks. We may therefore ask a general question about all of the sources used in answering a query, or we may ask about a specific part of the answer (e.g. where did you find the name "Asakusa View"). We refer to the issue of *what* as granularity.

The question of *where* information comes from actually takes on several dimensions in the context of evaluating data quality. Broadly speaking, a user might wish to know the publisher or source of information as a heuristic for judging the reliability of specific facts. Perhaps more significant, particularly in the context of the World Wide Web where reuse and redistribution of data is standard practice, is the question of where one particular data source received its information. As is the case with integration, data transmitted through several layers of redistribution often may suffer from successive filtering or translation, whether intentional or not (Lanter 1991; Woodruff and Stonebraker 1997).

Knowing from *where* a specific piece of information derives is useful for assessing the veracity of a specific data item. However, evaluating the quality of an answer with respect to the question raises a second dimension of *where*. Knowing *where* an integrator or a user looked is useful for gauging the completeness of a particular answer. The information conveyed by one travel guide on lodging in Tokyo may be 100% accurate, but because it only lists hotels in the financial district, the quality of the answer with respect to the query is quite different.

Questions of data quality also raise the question of *when* data is retrieved. Certainly a user can document the date and time on which they pose a particular query and receive a response. However, knowing when a query is posed and a response is given addresses only one dimension of *when*.

Related to *where*, the user might like to know *when* the data source last updated its information. For example, over what period of time is data archived or how frequently is data updated? As discussed below, some data sources preload data into distributed servers to enhance performance. As a result, however, data quality may suffer. Recall that the

(reduced) quality of cached data was at the heart of one of eBay's complaints against BE (eBay v. Bidder's Edge 2000).

Quality, of course, is only one motivation for a user's interest in attribution. Finding additional information is a second reason users might wish to know the attribution of data. The issue of search raises some additional dimensions to the question of *where*. Whether for assessing quality or for finding additional information, a user might generically ask *where* did the integrator look for the answer. In the same way that a user might wish to know about the veracity of a specific item of data, one might search for information related to a specific item of interest in the original answer. This was our original issue of *what*. General interest in the entire query answer is referred to as *coarse grained* result granularity. *Fine grained* result granules focus on specific values in the answer.

Just as a result has varying degrees of granularity, so to do sources. For example, knowing that information came from the public library is perhaps accurate but less useful than knowing a particular reference text. Moreover, consider the issue of Web navigation. Some sites are quite complex and tedious. The concept of "deep linking," which we will refer to below in the context of Ticketmaster, will introduce more about the concept of source granularity. Deep linking also has relevance outside the context of the Web. The difference between a reference list and a footnote illustrates the difference between coarse and fine grained source references.

We began defining the problem space by revisiting user interests in attribution. We now turn to the question of data integration to raise general data provider interests in the same issue. To understand how user and provider interests relate with respect to attribution, we begin with a definition of integration.

1.3.2 What is integration

To extend our understanding of attribution, we offer a stylized description of a prototypical integrator. We expand that definition into a taxonomy of different functional architectures for integration. The taxonomy allows us to systematically identify additional attribution challenges.

As expressed in the example of Chapter 1, the aim behind integration is to provide users with a single, uniform interface from which they can access heterogeneous, distributed data in a transparent fashion (Chawathe et al. 1994; Goh 1997; Levy, Rajaraman, and Ordille 1996; Quass et al. 1996). As illustrated in Figure 1.2, users pose queries to the integrator as though the integrator were a single, monolithic data source. Note that the data used to respond to the query could come from one or more underlying sources. The integrator might manage data of its own in addition to content from external sources. External data might be fetched in real-time, cached from previous queries, or pre-fetched into a warehouse. External sources to populate the local cache or warehouse could include everything from Web sources and networked databases to warehouses or even other integrators.

For our purposes, integration strategies vary on three axes: value-added, data timeliness, and user scale. The first axis along which integrators vary is the degree of value-added that they contribute to the information that they collect from other sources. Some integrators are themselves data producers who collect data of their own while the opposite extreme constitutes actors who merely act as a conduit for data from external sources. Along this continuum, integrators provide various value-added services including context integration to resolve semantic differences between data (e.g. reconcile hotel prices listed in Japanese Yen, US Dollars, Swiss Francs, etc.) (Bressan et al. 2000; Goh 1997; Goh et al. 1999) and de-duplication (e.g. merge listings so that the same hotel is not listed multiple times from different sources).

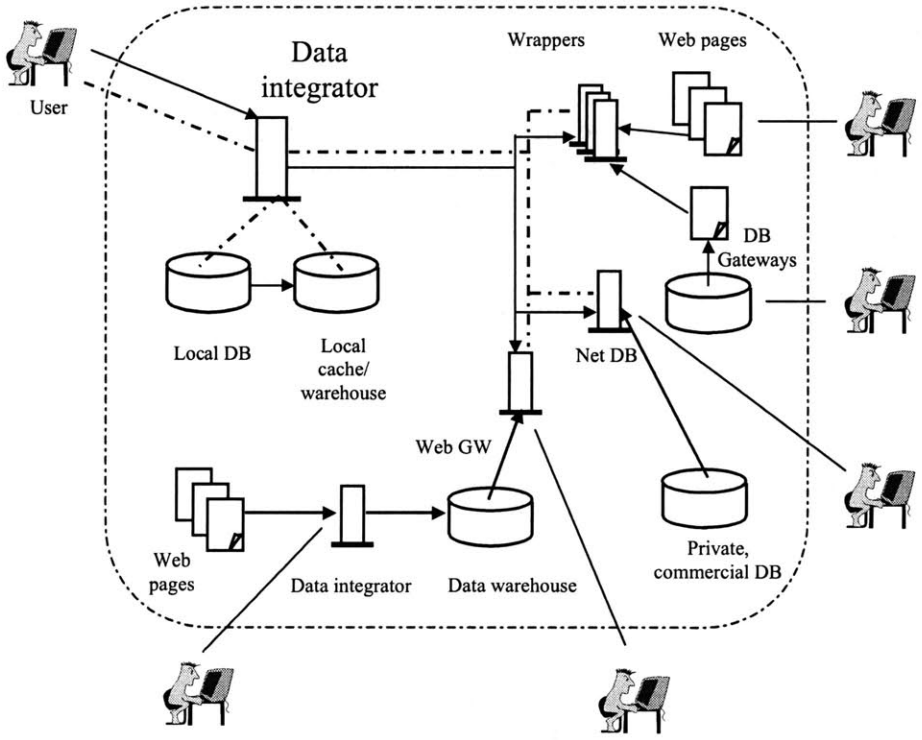

Figure 1.2 Integration architecture

Timeliness defines a second axis. Real-time queries are one extreme of data timeliness. In a real-time query, the integrator accepts a user query, submits a corresponding query to underlying sources, and provides an answer the instant the integrator receives the data from the external sources. BookFinder.com, for example, submits real-time user requests to services covering over 20,000 sellers of new, used, rare, and out-of-print books (BookFinder.com). Delays due to server load, network congestion, etc. however, are only magnified by real-time query integrators; such delay can prove costly. Zona Research estimates that total e.commerce losses due to user frustration with unacceptable download times exceed US$4.35 billion per year (Wong 1999). Archiving strategies such as caching

and warehousing contrast real-time services. These alternatives not only improve performance by pre-fetching but also facilitate the incorporation of value-added services. The penalty is data timeliness. Users may end up receiving data that is already outdated (eBay v. Bidder's Edge 2000; Kaplan 2000).[1] Strategies such as caching only query results rather than anticipating and pre-fetching or using time-to-live variables fall along this continuum.

A final axis is the degree to which integrators aggregate user requests to capture economies of scale in query processing. Some services process queries and populate caches in response to specific user requests. Others, such as those who pre-fetch, effectively amortize the cost of a single, external request over a population of users. A nuance on scale economies is management not only of queries but also the cache. So that multiple users could benefit from a single cache update, all users might share and access a single cache. At the opposite extreme, an integrator could maintain a separate cache file for every user.

We depict the relationships between these axes in Figure 1.3. As before, we use the spheres to place certain examples in the multi-dimensional space for illustrative purposes. BookFinder was an on-line book merchant. In response to a specific user's title search, BookFinder would invoke a real-time query to identify prices at competing on-line book sellers (e.g. Barnes & Noble bn.com) and then undercut the competing price (Bailey 1998). BookFinder was integrating data on behalf of a *single user*, in *real-time*, and providing *value-added* by way of price comparisons. We might think of mySimon as providing a similar value-added service. However, mySimon preloads product and price data from external merchants in anticipation of future requests rather than in response to specific requests. mySimon therefore *warehouses* data on behalf of *multiple users* to provide the *value-added* service of comparison shopping. Sites that list real-time stock prices, by contrast, provide a generic (meaning that it is available to *multiple users*), *real-time* service with *little value-added.* Any number of sites list real-time stock prices.

1.3.3 Provider interests

Reviewing different types of integration services helps to clarify the interests of different providers as opposed to the interests of users. To begin with, providers have a similar interest in *what* information is taken from *where* and *when*. Any single provider plays the role of a source from *where* a user collects data. Intellectual property considerations directly raise the question of *what* information is taken from individual sources.

As noted in our taxonomy of integration, the values of different types of data vary according to time (some content might even move into the public domain). Therefore, knowing *when* different pieces of information (*what*) are taken can also prove significant.

---

[1] Interestingly, in some cases, such as stock quotes, delay is a way of differentiating users. See Hoovers.com, eSchwab.com, etc.

Providers, however, are interested in more than just *what, where,* and *when*. Intellectual property concerns additionally ask *who* is taking information, *why* the information is taken, and *how* the resulting content is used. "*Who* takes the content" addresses the straightforward question of who should pay for the content that is taken. However, the issue can prove more subtle, particularly in the context of integration.

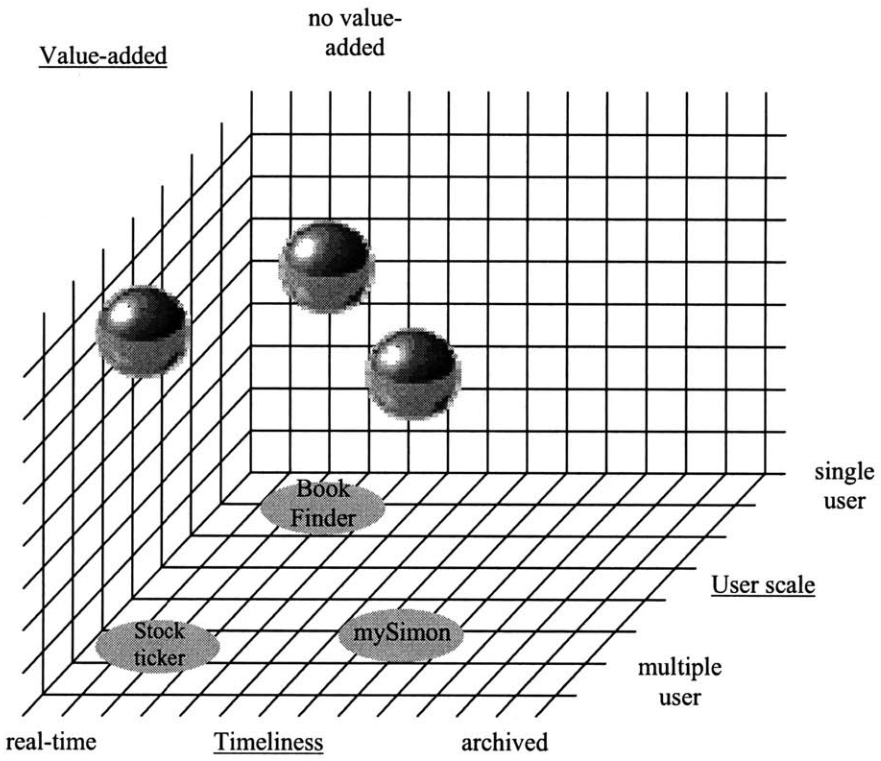

Figure 1.3 Integration strategies

Consider first the observation that an individual user might represent more than just herself. (For our purposes, we will reference this issue as the question of *why* information is taken.) Data integration services that collect content into a shared cache (irrespective of whether the data is pre-fetched or gathered in response to an initial query) exemplify individuals that represent or "stand in the shoes" of a community.[2] Likewise, a user of our hypothetical travel information integration service might be collecting hotel lists for a group tour.

[2] Consider also the interesting role of software-based infrastructure services (a.k.a. Content Delivery Networks (CDNs)), such as Akamai, that mirror and distribute data for balancing network traffic. Infrastructure services are outside the scope of this thesis (see Chapter 7). However, in general, we observe that CDNs use attribution data as indexes into distributed caches for constructing dynamic pages in response to client requests on application services (Akamai 2001).

Complementing the question of *why* is the question of *how* the content is used. Individual end users are, by definition, those who do not redistribute; use is limited to a single individual. Integration, however, is defined by reuse and redistribution. In integration, recall that user scale may vary from redistribution for single individuals (perhaps in answering a query by aggregating data gathered from multiple sources) to an auction aggregation service like Bidder's Edge that serves a broad population base. By the same token, content, once taken, may be used as-is or instead incorporated into some other, value-added products and services. Redistribution that competes directly with the original content provider raises different intellectual property considerations from reuse in value-added products and services that serve highly differentiated, niche markets.

We elaborate upon constituencies and their respective interests in our Policy Analysis. However, an overview of integration and its stakeholders provides a sufficient framework for defining the attribution problem space.

1.3.4 The attribution problem space for data integration

The attribution problem space, shown in Figure 1.4, that emerges from our taxonomy of integrators closely follows the dimensions along which integrators vary. We borrow from Lasswell (1948) to summarize the problem space in terms of *who*, *what*, *where*, *when*, *why*, and *how*. W*hat* and *where* correspond to our initial intuition behind attribution of "where does it come from?" Combined with *when* and *why*, the four concepts correspond to the axes that describe integration architectures while *who* and *how* address the relationship between different stakeholders in the attribution problem space.

With respect to a given query, *who* posed the query? Was it an end user or an agent representing a user? Was the query posed directly to some underlying data source or to an integrator? *What* information did the integrator use to answer a specific query, and from *where* did the user collect each piece of information? Some of that information might have been locally generated while other content might have come from a local cache of remote content. *When* was the specific request processed? Was content to answer the query gathered in real-time, or was any information collected from a local cache or data warehouse? *Why* did the user (perhaps an integrator rather than an individual) process a specific remote request? Did the user execute independent requests for herself, or perhaps serve as a representative, aggregating the query over a number of users making the same request? Finally, *how* did the integrator use the different pieces of information that were collected from external sources? Did the integrator clean, update, reformat, or otherwise add value to the data? Did the integrator take data to compete directly with a source or perhaps apply data from one domain to a completely different market?

1.3.5 Motivation for attribution

We have identified a number of challenges that help define what we mean by attribution. These distinctions are not merely academic. In the context of user and provider interests, we saw that three general motives for attribution are: verifying data quality, searching for related

information, and intellectual property. Looking carefully, we can see that each distinction that we drew has specific bearing on one or more of these motivations.

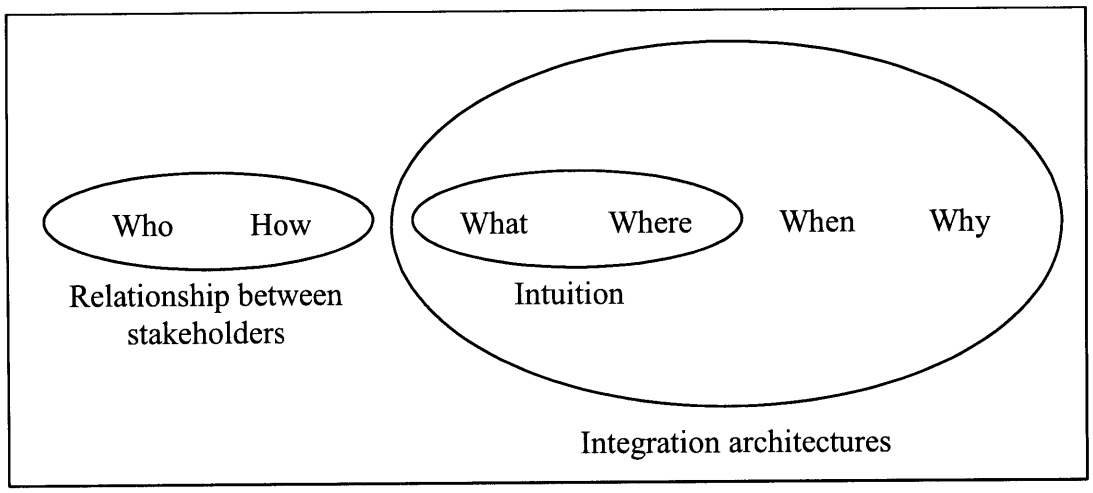

Figure 1.4 The attribution problem space

1.3.5.1 Data quality

Knowing the sources that provided individual answers in a query result helps vouch for the accuracy or correctness of a specific fact. For example, do we believe that a hotel name is spelled correctly or that the prices listed for a specific hotel are current? If the source is reputable, we are much more likely to accept the accuracy of the spellings or facts.

Knowing all of the sources explored to answer a query and whether any answers were found there speaks to the comprehensiveness or completeness of a query result. For specific answers, knowing the sources used in each step of the query process helps verify the accuracy of the answer set. For example, whether a hotel is close to a national landmark is not a function of whether the hotel's name is spelled correctly. The correctness of the answer depends upon whether the sources used to evaluate query conditions, such as the regions in which hotels and national landmarks are located, are accurate and up to date.

Finally, knowing whether there are multiple ways of deriving an answer, multiple sources for a specific value, or whether there are contradictory results are all ways of reinforcing (or diminishing) confidence in a specific value or answer.

1.3.5.2 Search

Once we have a list of hotels that satisfies our criteria, we might want to read more about a particular hotel or tourist attraction. Identifying the specific source or guide that mentioned the hotel or suggested a site is one heuristic for finding relevant, additional information.

As an analog to the quality of the answer to our query, we might want to read from sources that provided contradictory results or answers used in evaluating query constraints. Doing so could help answer questions like why certain answers we might otherwise have expected were excluded. For example, Mount Fuji is nowhere near Tokyo.

Finally, if we wanted to share our information with colleagues who might similarly be planning vacations, we could either share with them our search results or instead, share with them our search strategy. They could apply our strategy to other destinations or refine the strategy to suit their own tastes. Moreover, by identifying multiple strategies for finding the same answers, we can identify the critical or important sources as those which are common to more than one of our derivations.

1.3.5.3 Intellectual property

Distinctions in attribution are similarly important for determining who to acknowledge or who to compensate (e.g. through micropayments) for the results of a specific search. One policy might compensate only those sources that provided an answer. This might be akin to browsing a mall but purchasing only what satisfied the consumer's needs. However, from a different perspective, an answer, in its completeness embodies not only the values included but also those that are excluded. Consequently, perhaps every source used in evaluating a query should be acknowledged.

Granularity has specific relevance to the assignment of attribution for intellectual property purposes. In the print world, the difference in precision between a bibliographic entry and a citation is well established. This difference corresponds to our notion of source granularity. Similarly, works referenced or cited might be aggregated over a single chapter or an entire volume. This corresponds to our attribution characteristic of result granularity.

In the distribution and redistribution of on-line data, similar distinctions apply. Ticketmaster Online-Citysearch, Inc. (TMCS), for example, partners with Zagat.com to provide restaurant listings and reviews for major cities in the United States. However, rather than attributing every restaurant listing or even the sections on restaurant listings, TMCS simply lists Zagat.com as a national content partner. Coarse granularity is clearly not acceptable in all instances, however. The granularity of attribution, not the presence or absence of attribution, was central to the dispute between mySimon, Inc. and Priceman. In commenting on the case, the founder of Priceman "conceded that unlike many other meta-search engines, his site did not attribute specific results to the site that provided them. He maintained, however, that a sub-page on his site listed the seven or eight sites searched, and that mySimon was listed there (Kaplan 1999)." Why some services (e.g., Zagats.com and TMCS) might be content with coarse-grained attribution while others might not (e.g., mySimon, Inc.) is beyond the scope of our effort to define attribution but will be discussed as part of the broader attribution problem space later.

1.4 Summary

We introduced the concept of attribution with a simple example to both describe the problem and motivate the problem's significance. We then parameterized the problem space with a set of questions and related those parameters back to our original motivations for addressing the problem. These parameters will also serve a set of desiderata by which we may compare different approaches to the issue of attribution. To comprehensively address the problem, an attribution strategy should identify:

Who is querying the data. The question of *who* is further qualified by *why* and *how*.
Why the query is posed (i.e. is this for a single user or as a proxy for many others); *how* the information is used (i.e. for personal use, to develop value-added products, in competition with the data source, etc.)

Attribution must also address *what* information is being sought and *where* each individual data item comes from. The relationship between *what* and *where* is further qualified by the issues of multiple derivations and granularity. The same content (i.e. *what*) may come from different places (e.g. redundant sources or multiple derivations). We may also specify the relationship at varying levels of detail (i.e. granularity of *what* and *where*). In the context of print, we might compare bibliographies (coarse grained) to footnotes (fine grained).

Finally, consider the question of *when* content is taken. Depending upon the user's purpose, *when* may significantly affect quality. Conversely, if old enough, *what* is taken and *how* that content is ultimately used may not matter.

2 Related work

As evidenced by the history of research in citations and references, attribution existed as a general principle of data management long before the advent of digital media and electronic databases (IFLA 2002). The need for attribution is only exacerbated by the medium for widespread data reuse and redistribution that defines the World Wide Web. Therefore, it is perhaps not surprising that there is a great deal of research that relates in one measure or another to the attribution problem space as articulated in Section 1.

While the formal and pragmatic technologies reviewed in the thesis excerpt on technology perspectives addresses the relationship between *what* and *where*, the attribution problem space itself is much broader. To more completely address the problem space in its entirety, we expand the scope of the research presented in our technology perspectives to consider policy alternatives as well. As is the case for technology alternatives, the breadth of the problem space encompasses a wide range of related work. In this section, we focus in particular on policy approaches similar to our own.

Much of the research literature on policy perspectives to the attribution problem space is a response either to specific policy proposals or to related legal proceedings (e.g. eBay v. Bidder's Edge referenced in Section 1). As a consequence, we begin our survey of related policy work by examining recent policy proposals. We then consider some of the academic literature addressing the same topic. Because legal proceedings focus on the existing regime we reserve that discussion for Section 3. In Section 3, we provide a comprehensive review of the status quo policy approach to questions of *who*, *what*, *where*, *when*, *why* and *how*.

2.1 Recent policy proposals

The role that property protection plays in quality, remuneration, and search, the motivations cited in Chapter 1, is reflected in the comments of librarian Ingrid Shaffer: "Few notice who provides the data or who pays for it. But we should, because the issue affects its quality and availability ... Without better government copyright protection, where is the incentive for such businesses to provide high-quality information? (CADP 2000)."

Passage of the European Database Directive (EDD) in 1996, which requires reciprocal U.S. legislation in order for U.S. products to receive equivalent protection in Europe (Hunsucker 1997), brought the need for a coherent U.S. policy into sharp relief. Since that time, the U.S. policy approach to the attribution problem space has centered on intellectual property. In part spurred by European action, Representative Moorhead introduced H.R.3531, the Database Investment and Intellectual Property Antipiracy Act, in May of 1996. The legislative history since that time has included H.R.2652 introduced in 1997 by Representative Coble, S.2291 introduced in 1998 by Senator Grams, H.R.354 introduced in 1999 also by Representative Coble, and H.R.1959 introduced in 1999 by Representative Bliley.

Every Congress from 1996 through 2000 has considered attribution related legislation for data reuse and redistribution. The absence of explicit U.S. policy only magnifies the significance

of action in other nations. The combination of domestic pressure and international action suggests that U.S. policy is more a question of when rather than if. We therefore review the two most recent policy proposals from the perspective of the attribution problem space as exemplary of current policy alternatives. A brief overview of the EDD, in the context of the attribution problem space, is provided for contrast.

H.R.354, The Collections of Information Antipiracy Act, is the third such legislative proposal to bear that title in the past three years. The *who* in the attribution problem space is answered in H.R.354 as any consumer of a commercial database product. No explicit mention is made of proxies who might gather data on behalf of a client therefore *why* is unaddressed in the problem space. Although defined ambiguously, H.R.354 prohibits the taking of "all" or a "substantial part" of a commercial product in a way that would cause material harm to the primary market or related markets for the original database. The restriction applies for fifteen years. Consumers and competitors are free to gather the underlying data from the original sources at any time.

What from an attribution perspective is thus defined as "all" or a "substantial part." Of greater significance is the question of *how* the content may be used. Subject to fair use permissions articulated for science, education, and personal use modeled on the Copyright Act, any use that might cause material harm in both primary and related markets is prohibited. Proponents of H.R.354 argue that strong property rights are necessary in order to incent initial data gathering (Aber 1998; Corlin 1998; Garland 1999; McDermott 1999; Tyson and Sherry 1997; Winokur 1999; Zuckerman and Buckman 1999). Opponents argue that such limitations threaten to curtail legitimate science and education as well as stifling innovative data reuse (Hammack 1998; Lederberg 1999; Linn 2000; Neal 1999; Phelps 1999; Reichman and Samuelson 1997; Reichman and Uhlir 1999; Samuelson 1992). Reconciling these positions is reviewed in greater detail as part of the Policy Formulation exercise in Chapter 8.

The attribution technologies addressed earlier address the relationship between *what* and *where*. H.R.354 answers the question by noting that prohibitions apply to the data collections gathered by a particular producer. H.R.354 does not prevent users from accessing and (re)gathering the data from the original data sources. Likewise, H.R.354 explicitly establishes an upper bound on *when* users may take data. After 15 years, property protections on a collection cease to apply. Whether data maintenance and quality checking warrant renewal resulting in perpetual protection is an open question and beyond the scope of this research (Reichman and Samuelson 1997; Tyson and Sherry 1997).

Contrasting the strong property right proposed by H.R.354 is the Consumer and Investor Access to Information Act introduced by Representative Bliley as H.R.1858. *Who* and *why* are defined as in H.R.354. Again no mention is made of proxies who gather data on behalf of individual users. H.R.1858 prohibits duplicating or copying to create a collection of "substantial similarity" to an original commercial database product. More specifically, copies are prohibited from sale or distribution in competition with the original provider. The explicit intention is to prevent the displacement of sales or licenses that would threaten a rights

holder's recovery of the initial data collection investment. Therefore, a fixed time limit on the duration of the right is not established. The restriction ambiguously extends only to recovery of the original investment. Of course, as before, consumers and competitors are free to gather the underlying data from the original sources at any time.

For all of the ambiguity in both legislative proposals, H.R.1858 is considered much less restrictive than H.R.354. With respect to the attribution problem space, H.R.1858 defines *what* may be taken in terms of outright duplication. Moreover, restrictions on *how* one may reuse data are more limited. The language explicitly acknowledges the need to protect a data gatherer's initial investment in collecting, but focuses primarily on ensuring public access to the resulting collection. The classic intellectual property tradeoff between private investment and public access is explored as a part of the Policy Formulation exercise in Chapter 8.

As in H.R.354, users are always free to gather data themselves from the original sources, freeing them from any additional restrictions. H.R.1858 therefore only applies depending upon *where* a user gathers or duplicates data from. However, no fixed time limit is set on the duration of this restriction. There is no bound on *when* data collections enter the public domain. Instead, the legislative history surrounding the bill focuses again on the tradeoff between private investment and public access. The implication is that protection should extend no longer than the time required to recover investment; the assumption is that investment recovery will take far less time then existing, statutory provisions for intellectual property such as copyrights or patents (databasedata.org 1999a; b).

The EDD, which magnified the existing U.S. policy interest in database legislation, is directed at the questions of *what*, *why* and *how*. Specifically, database producers are granted the "(1) right to prohibit the extraction of, and (2) the right to prohibit reutilization of all or a substantial part of the database contents."[3] The EDD does not draw distinctions between end users and intermediaries; in so doing, the EDD does not concern itself with *who* extracts content. Instead, focus is placed on "reutilization." In the context of the attribution problem space, we might think of "reutilization" as the intersection of *why* and *how*. We use *why* to categorize users (or software agents) that extract data on behalf of one (or more) users. Similarly, the attribution problem space defines *how* to document whether data is reused in direct competition with the initial producer. Untested in the European courts, there is no interpretation of how broadly the initial database producer may constrain *why* or *how* under the EDD. Moreover, rights conferred by the EDD are renewable in the production investment. Consequently, periodic investments that are proportional to the initial creation investment and made for the purpose of updating database contents could conceivably extend the right indefinitely (Nissen and Barber 1996). Under an interpretation that permits perpetual renewal, delaying or time-shifting data reuse (i.e. the attribution dimension of *when*), whether by caching or otherwise, provides no relief. Further discussion of stakeholder interests in and the implications of policy measures like the EDD is deferred to the policy analysis and formulation in Chapters 7 and 8.

---

[3] EDD art 8(2) J.L. 77/20 at 26 in (Hunsucker 1997)

In Chapter 8, we develop a policy proposal that assumes the Constitutional mandate of "progress of science and the useful arts" (U.S. Constitution Article 1 Section 8) as its primary goal and builds on two theoretical frameworks for intellectual property, game theory and entitlement theory (Calabresi and Melamed 1972; Gibbons 1992). As a consequence, we propose a liability approach that focuses heavily on *how* content is used and less on *what* is taken or even *when*. We explain in Chapter 7 how the question of "*why* content is taken (meaning on whose behalf)" may crucially affect the market model by which a vendor anticipates recovering their investment. Our policy proposal thus also incorporates consideration of *why*. We concede the possible role that a statutorily determined time frame governing *when* may be appropriate. Following H.R.1858, we accept that the issue may be important, but leave an analysis of optimal protection duration for another investigation.

2.2 Related academic literature

There is a large body of academic literature related to the policy focus of this thesis research. Much of the existing work, however, is either in direct response to current interpretations of status quo policies (i.e. Court rulings related to database (re)use) or research in the broad space of intellectual property, without any specific emphasis on information technologies and the attribution problem space. While we will refer to existing work throughout Chapters 7 and 8, we focus here on new policy approaches addressing the attribution problem space. Given this limitation, relevant work is divisible into policy approaches to rights in data specifically and information technology in general.

2.2.1 Related work on database rights

Research directly addressing rights in data have tended to derive from two differing intellectual property foundations. The first foundation regards property rights in authorship as natural law. This Romantic approach to intellectual property rights has its greatest following in the European intellectual property tradition (Merges et al. 1997). By contrast, the U.S. Constitution establishes intellectual property as a balance between public access to information and private incentives to gather or produce said content (Merges et al. 1997). Ginsburg (1990) compares and contrasts the two positions with respect to property rights in data. She concludes that works of "low authorship," such as collections of facts, appropriately fall between the need for strong regulatory protection and no protection whatsoever. Accordingly, she proposes compulsory licensing as a middle ground between intellectual property monopolies that could discourage innovative reuse and zero liability, which would destroy any incentive to produce.

Patterson (1992), in arguing from a natural law framework, also categorizes collections of facts as works of "low authorship." Rather than borrowing from the intellectual property regime, however, Patterson turns to trade regulation. He argues that a Federal statute in unfair competition is the most appropriate means for supporting both educational and scientific interests in access to data and protecting the broader public interest in access to information.

Public access to information as captured in First Amendment principles (U.S. Constitution) is the foundation from which Pollack (1999) makes her argument. By setting out free flow of information as the paramount objective, Pollack concludes that broad restrictions on data reuse, such as that proposed in H.R.354, constitute an un-Constitutional prior restraint on speech, irrespective of whether the speech (the database) is commercial. Pollack follows Patterson's consideration of trade principles and concludes that the Court's decision in *INS*[4] is flawed. Limited reuse with appropriate remuneration that does not compromise the original producer's ability to recover their costs (displace sales) is appropriate. Policy must balance the twin Constitutional free speech and intellectual property provisions.

Reichman and Samuelson (1997) likewise build from a Constitutional perspective. They evaluate policies based upon those which would best promote science and education in general but also address innovative data reuse. Theirs is a comprehensive work that surveys status quo policy through time of publication (i.e. legislative proposals through the European Database Directive and leading to H.R.3531). They advance an intellectual property-based, modified liability approach to balance producer and consumer interests.[5]

The National Research Council (NRC), in their report Bits of Power, considers the problem of data reuse and redistribution (NRC 1997). Together with a subsequent report The Digital Dilemma (NRC 2000), the NRC reviews both technology and policy alternatives for addressing the attribution problem space overall. Reflecting their Federal commission, the NRC reports focus on scientific and educational interests in data reuse. Unlike the other scholarly work referenced above, however, the NRC reports relates the legal principles to one set of underlying economic principles, transactions cost economics. From this foundation, the NRC supports policies with exceptions for science and education as well as additional research into the economics of the database industry to better understand policy impacts.

Tyson and Sherry (Tyson and Sherry 1997) adopt a similar, economic foundation. They develop the framework for categorizing data which we adapt in Chapter 1. From that basis, they review the state of the industry and conclude that Federal intervention, through a strong property right in data, is necessary to ensure a vibrant market in database creation. Fears about monopolization and market power are answered by a competitive marketplace. Protecting databases, they argue, does not preclude equal access to equivalent base sources, either because the raw data remains in the public domain or because anti-trust legislation would restrain sole-source providers.

With the exception of the NRC reports, none of the literature addressing data rights in particular captures the full scope of the attribution problem space. Like Reichman and Samuelson (1997), we begin from the premise that the principle objective of intellectual property legislation is the promotion of science and the useful arts. Like the NRC, we rely

---

[4] We summarize and elaborate on *INS* in particular and misappropriation as doctrine in Chapter 7.

[5] In (Reichman and Samuelson 1997) unfair competition is also presented as a policy alternative. However, they conclude that a modified liability intellectual property rule rather than unfair competition, rooted in trade regulation, is preferred.

upon economic frameworks rooted in transactions cost economics (Milgrom and Roberts 1992). While we categorize data in a manner that follows Tyson and Sherry, we decompose the industry into different market models in Chapter 7. Those market models, combined with a review of the science in database creation in Chapter 8, lead us to a different set of conclusions. It is this combination of both technologies and economics underlying the industry, as well as technology and policy alternatives for protection, that makes this analysis unique.

2.2.2 Related work on IT and IP

In addition to research addressing databases directly, there is also a more general body of literature on intellectual property and information technologies from which we borrow. Perritt (1996), Hardy (1995; 1996), and Merges (1994; 1996) all build from a transactions cost framework. They consider the impact of information technologies on various transactions costs associated with bargaining for intellectual property. Policy proposals are differentiated based upon the cost of bargaining according to the Entitlement framework articulated by Calabresi and Melamed (1972).

Hardy focuses on the promise held by information technologies for decreasing the costs of intellectual property transactions. In particular, he focuses on three costs. First, IT dramatically decreases search costs, the cost of identifying products and parties with whom to transact. Second, IT supports the ability to define and enforce property boundaries. Referencing technologies, such as the attribution defined in Part 1 of this thesis, support property claims. Technologies such as encryption and access controls enforce those property claims. As a consequence, Hardy concludes that strong property rights are warranted.

Using the same framework, however, Perritt comes to a different conclusion. Perritt defines cost models for production and piracy of digital content, respectively. In so doing, Perritt first points out that in many respects, the costs of production and piracy are not so divergent. Moreover, once itemized, he observes that appropriation is not necessarily costless. He concedes that digital copies are inexpensive to both produce and distribute. However, the concomitant decrease in enforcement costs through monitoring and access controls suggests to Perritt that perhaps the status quo is adequate. Combined with contracts, Perritt argues that the status quo policy alternatives for managing intellectual property in the face of new IT is adequate.

Merges begins with the same foundation. Rather than fitting the case for intellectual property into the Entitlement framework as originally presented by Calabresi and Melamed, however, Merges extends the framework. He argues that some forms of new information technologies alter the economics sufficiently to expand the liability property dichotomy to a third classification, private liability rules. As exemplified by collective rights agencies, Merges argues that private liability rules are best suited to addressing certain categories of intellectual property. "Private" liability rules, by definition, are not a government policy. However, Merges does suggest that government sponsored research into enforcement and monitoring

technologies as well as the creation of strong property rules may incent the creation of the private institutions that establish private liability rules.

3 Policy analysis

In Part 1 of this thesis, we introduced a theory of attribution as a technology for relieving some of the tension that arises from the emergence of integration tools. However, the technology only provides a *means* for balancing user needs and provider incentives. Motivation to use the technology, whether by legislative or market mandate, is unresolved. Therefore, in the next two chapters, we adopt a broader, policy perspective on integration and attribution-related problems.

Chapter 7 is a policy analysis. We survey the current policy landscape in terms of the problem space as defined in Chapter 1. We then review the status quo legal framework from the perspective of the Chapter 1 problem parameters and look more closely at the stakeholders, and their respective interests. To close, we intersect the existing policy framework with stakeholder interests to arrive at a consensus on the need for, if not the nature of, change.

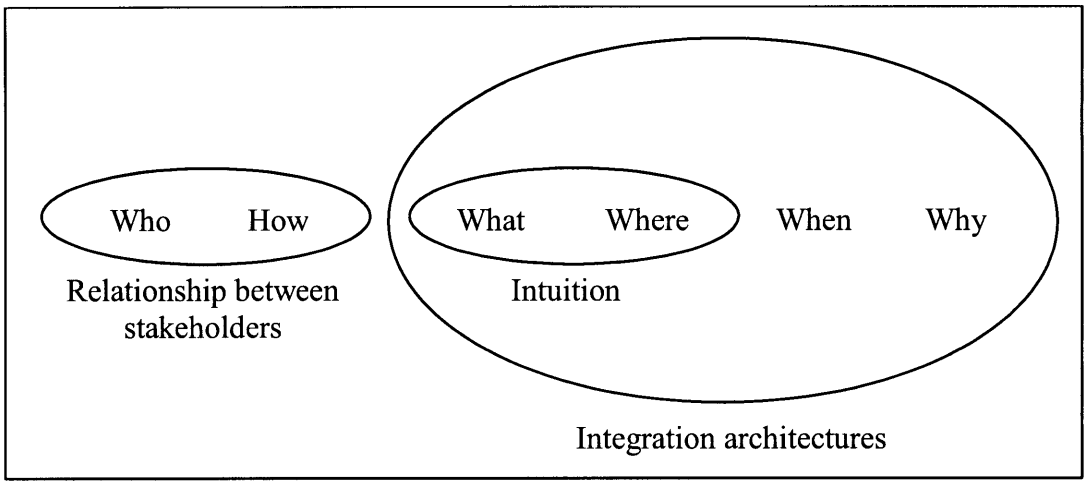

Figure 3.1 Attribution problem space (redux)

3.1 Defining the problem space: integration challenges, redux

In Chapter 2, we established the problem space by considering the process of integration and stakeholders in the process. As a consequence, we concluded that we could describe the problem space in terms of the following integration dimensions: *what* is taken, *where* the data comes from, *why* (on behalf of whom) and *when* is the content taken. To integration we added the stakeholder relationships *who* is taking and *how* is the content used.

We use this same framework as a vehicle for structuring our tour of the policy landscape. Because the framework describes the problem space, we note in advance that particular policies may ultimately address multiple dimensions at the same time.

3.2 Surveying the status quo

The United States has a history of intellectual property protection that dates back to the Constitutional framers' Congressional mandate to "promote the progress of science and the useful arts (U.S. Constitution1787 at Art 1. Sec. 8)." The need for intervention was anticipated to balance incentives to create with a public interest in dissemination (Drahos 1996; Merges et al. 1997). The problem of data reuse and redistribution, though exacerbated by information technologies, has already existed for some time. In this section, we describe how the existing policy framework covers the attribution problem space. We consider, in turn, policies that affect *what* can be taken and limitations that stem from requirements on *where*. With an eye on integration, we then ask about constraints on *who* may take and *why* (upon whose behalf); finally we ask *how* content may be reused and *when* content is appropriable.

3.2.1 What

The anchor of prevailing database policy protections was crafted by the Supreme Court in its ruling *Feist v. Rural Telephone Service Co., Inc.* (Feist v. Rural 1991). Rural Telephone is a public utility that publishes a white pages listing of all its customers. Feist sought to publish a regional phone book combining the customers serviced by Rural Telephone with those in a number of surrounding service areas. Feist sought a license from all of the concerned utilities for use of their respective customer listings. Of all the utilities in question, only Rural Telephone refused. Rather than produce a non-comprehensive directory, Feist copied the Rural Telephone listings without authorization. Rural Telephone subsequently sued under copyright, claiming ownership of the listings copied by Feist.

The Supreme Court ruled in favor of Feist, articulating the position on databases and copyright that persists today. First, the Court rejected the notion of a copyright in facts (the contents of the database in question). The court observed that, while a database compiler might be the first to discover or publish some fact about the world, the database compiler in no way "creates" the fact. Second, the Court conceded that a copyright could exist in a creative selection or arrangement of facts. An exhaustive, alphabetical listing of all customers fails this standard. Third, by extension, the Court rejected the notion of a copyright based solely upon "sweat" or material investment in the collection (Samuelson 1992). Hard work does not, in and of itself, justify intellectual property protection. Therefore, all else being equal, the decision in *Feist* governs *what*. A third party has the right to extract, reuse, and re-distribute *what*: the facts collected and ordered into a database by another.

Though database intellectual property was already an issue because of the growing threat from data networks and the Internet, the clamor raised by database producers following *Feist* led to discussions by the World Intellectual Property Organization (WIPO) and subsequent adoption of the European Database Directive (EDD) in 1996. Often confused with privacy legislation passed that same year governing the collection and use of identifiable consumer information, the EDD most notably establishes a renewable property right in an ordered collection of facts based upon the investment required to collect and organize these facts

(Hunsucker 1997). Specifically, producers are granted the "(1) right to prohibit the extraction of, and (2) the right to prohibit reutilization of all or a substantial part of the database contents."[6] The British law firm of Harbottle & Lewis notes that the right exists in the database as a whole and not in the individual facts, which could be recreated by a second comer without penalty, though often at considerable expense (Nissen and Barber 1996). Moreover, the right is renewable in the investment. Therefore, periodic investments that are proportional to the initial creation investment and made for the purpose of updating database contents could extend the right indefinitely. Finally, the EDD includes a reciprocity clause denying equivalent protection to products from countries without equivalent protection (Bond 1996).[7]

As an alternative to statutory protection, companies such as Bloomberg or Lexis Nexis apply contracts to protect their content. The question of *what* may be reused with respect to databases is illustrated in *ProCD, Inc., v. Zeidenberg* (ProCD v. Zeidenberg 1996). Zeidenberg purchased multiple electronic databases of telephone listings from ProCD as well as the packaged software to query and access that data. After loading the data onto his own computer, Zeidenberg created custom software to search the aggregated data set. The directory, accessed through Zeidenberg's software, was then marketed on the Internet. ProCD sued, claiming violation of the licensing agreements contained inside the boxes of the products as well as embedded in ProCD's data access software. At issue were two key questions. First, was "use" a legitimate standard of assent with which to bind Zeidenberg to the terms of the shrink wrap license and second, were the contract binding, could it be used to preclude rights otherwise granted by Federal law (namely, the right to reuse facts as per *Feist*) (Elkin-Koren 1997)?

The trial court concluded first that the standard of assent was too low to form a binding contract and second that a valid contract could not preempt Federal copyright law. The appeals court disagreed on both counts. Were the standard of assent too low (e.g. use of the product constitutes assent), reasoned the trial court, such a contract would be meaningless. Everyone would effectively be subject to the contract. In disagreeing, the appeals court asserted instead the standard of substitutability. The contract would not affect the right to gather the same data independently, and the initial product could always be returned. The appeals court set a flexible guideline for subsequent contracts that is binding only on the parties (third-party integrators, in the analysis of this thesis) to the contract (O'Rourke 1997). Although contracts could also be written to address other dimensions of the problem space (e.g. *where*, *when*, or *why*), a number of other mechanisms discussed also apply.

3.2.2 Where

From our earlier description of integrators, we know that integrators might act as intermediaries, directing users to content stored and maintained by others. Alternatively

---

[6] EDD art 8(2) J.L. 77/20 at 26 in (Hunsucker 1997)

[7] Unlike the case law from which we may evaluate the boundaries of the U.S. policy landscape, the EDD has been largely untested in the courts. However, as noted in Chapter 2, the EDD has less to say on dimensions such as *who* and *where* and the limits on *when*, *why*, and *how* are uncertain.

integrators might cache content locally and draw responses to queries from the cache. Integrators who reference users to content stored and maintained by others use techniques similar to the HTML link common in today's Web. "An HTML link has two ends and a direction. The link starts at the 'source' end and points to the 'destination' end.... A link end [may refer] to some Web resource, such as an HTML document, an image, a video clip, a sound, a program, the current document, etc. (Raggett 2001)"

Integrators that answer the question of *where* by linking to external sources face at least two policy constraints. First is the performance copyright, and second is the question of trademark. Note that none of the copyright and trademark cases described in this subsection were decided in court. Every case was settled. Therefore, while no precedent exists, the cases outline how the existing policy infrastructure might be applied to the situations at hand.

In England, the on-line version of the Shetland Times sought relief from the on-line version of the Shetland News, a competing service. The News was providing a list of headlines that linked directly to the corresponding Times stories. By linking directly to the story rather than through the Times' front page, so called "deep linking," News readers bypassed the Times' banner advertising and missed the look-and-feel of the Times because the Times' frame was not similarly linked (Sableman 1999)[8]. Unlike caching systems, the News transported users to the Times site and users loaded the Times stories from the Times servers. A similar issue was raised in *Ticketmaster Corp. v. Microsoft Corp.* (Ticketmaster 1997) where Microsoft's Seattle Sidewalk city guide provided an up-to-date event list with deep-links to Ticketmaster's event listing and purchase page. At issue was denying Ticketmaster the user's click-through (Kuester and Nieves 1997).

In both instances, the integrator did not hide the fact that content came from an external source. The Times logo appeared by each story title (Sableman 1999). Sidewalk users viewed Tickemaster screens (Kuester and Nieves 1997). The challenge in these instances arose from the perceived loss of an author's "moral right" to control performance or, in the case of databases, patterns of access to content (Sableman 1999). The performance copyright, codified in the 1976 Copyright Act protects for creators, the right to perform a work. "The consequences follow from the feature of electronic communication that distinguishes it from the printing press: it is a process for performing, not publishing, works (Patterson 1992)." Deep-linking supplants the right of a host source to dictate a user's navigation, irrespective of *what* content is linked by *whom*, *when*, or *why*. As noted above, while both cases were settled out of court, the potential for devising such a policy instrument persists. Were such an argument to prove valid, not only would integrators be severely limited in *where* to process user queries, but also the traditional search engines, so popular on the Web today, could be found in gross violation.

---

[8] See *Shetland Times, Ltd. v. Wills,* F.S.R. 604, 1997 S.L.T. 669 (Outer House 24 October 1996) in (Sableman 1999).

In the *Shetland Times* and *Ticketmaster* cases, the potentially infringing integrators did include references to the Times and Ticketmaster. *Washington Post Company v. Total News Inc.* (Total News 1997) presents a more insidious example of what is possible by combining links with frames. Frames "allow authors to present documents in multiple views.... Multiple views offer designers a way to keep certain information visible, while other views are scrolled or replaced (Raggett 2001)." In linking articles from the Washington Post and other commercial services through a frame, Total News not only blocked advertising and identification from the originating sources, but also buried the originating URL within the Total News frame (Total News 1997). This means that Total News readers were not necessarily cued to the fact that specific stories came from external sources. Neither the frame, the banner, nor the URL in the browser window attributed particular stories.

Shetland Times, *Ticketmaster*, and *Total News* all highlight the potential for misrepresentation that stems directly from linking. Specifically, they introduced potential action under trademark violation. For *Ticketmaster* and *Total News*, under the Lanham Act, "the registrant of a trademark may obtain injunctive relief against any person who uses for commercial purposes a reproduction or imitation of a registered trademark, whether or not it appears on product wrappers.... (Effross 1998)." Specific issues include the threat of passing-off where a violator uses the trademark to unfairly associate a more obscure product with a better known brand and reverse-passing-off where a violator casts a better, competing product as his own. As previously stated, because all three cases were settled out of court, the validity of a trademark suit is still untested.

Complicating the trademark issues, which may unfairly associate one company's data with another company or product, is the danger of aggregating private data about individuals for which attribution can also raise privacy concerns. The analog to "passing-off" with products are the "false-light" and "right of publicity" standards for individually identifiable information. At issue is whether a link would associate a private individual with the aggregator in an impermissible fashion. Specifically, the association could cast the individual in a "false light" by causing others to believe that an association exists where one does not. Similarly the "right to publicity" standard argues that identifiable individuals should have the right to determine "who gets to do the publishing (Effross 1998)." Privacy issues are, however, beyond the scope of this thesis.

We have seen that trademark claims may affect integrators that link to external sources. The potential for misrepresentation is at least as applicable to integrators that retrieve content from external sources to integrate or otherwise add value. In Chapter One, we introduced the lawsuit brought by mySimon, Inc. against Priceman. Priceman was a meta-search comparison shopping service that integrated results from seven or eight different comparison services. mySimon's principal complaint stemmed from the determination that "although Priceman purported to search many price engines, in most instances it exclusively searched mySimon, usually without attributing those results to mySimon (Kaplan 1999)."

The problem, however, was less one of integration and more one of attribution. As noted by its president, "mySimon has no legal dispute with [other meta-search engines] because 'they don't take our results and strip away our name and branding and report those results as their own (Kaplan 1999).'" Priceman maintains that mySimon received attribution by virtue of mySimon's inclusion in a list of sites searched by Priceman. Priceman has since been shut down though the lawsuit persists. The lawsuit does not attempt to limit from *where* one may extract information. The lawsuit also does not suggest *how* the information may be used (e.g. in direct competition with mySimon). The lawsuit does assert that integrated content should receive attribution. Because a trial date has yet to be set, no court has had an opportunity to articulate what constitutes sufficient attribution.

3.2.3 Who

Though this thesis is interested in the (im)balance posed by integrators, the broader policy space recognizes that there are a variety of different users who query sources. As a status quo policy mechanism, trespass enables the selective regulation of *who* may take content from *where* regardless of *what*, *when*, or *why* that content is taken or *how* that content is used.

Recall from Chapter One the case between eBay and Bidder's Edge (BE). BE searches, extracts, and aggregates items and prices in a wholesale manner from a number of on-line auction houses into a single, comprehensive archive (Krebs 2000). User queries to BE are answered from the archive rather than directly from the underlying auction houses used to populate the archive. In *eBay Inc., v. Bidder's Edge*, (eBay v. Bidder's Edge 2000), eBay successfully won an injunction, pending the full trial in March 2001 (Kaplan 1999), enjoining Bidder's Edge from automatically extracting eBay listings into the BE, aggregate database.[9]

The judge's preliminary injunction was based on one of eBay's many claims. eBay successfully argued that the physical, computing resources that support an on-line host constitute chattels or any "species of property not real estate or freehold(Anderson 1893)." As chattels, the court concluded that, "Ebay's server and its capacity are personal property, and that BE's searches use a portion of this property (eBay v. Bidder's Edge 2000)." Actionable trespass of chattel occurs when unauthorized use results in damage to the owner (eBay v. Bidder's Edge 2000). Specifically, the judge agreed that automated searching by software agents might constitute trespass to chattels (Kaplan 2000; Krebs 2000). Trespass therefore becomes a viable means for regulating *where* an integrator might go to gather data for linking or caching.

However, actionable trespass has two elements. As laid out in *eBay*, the second part of the trespass policy instrument is to accept "the *possibility* that [the property owner] will suffer *irreparable harm* (eBay v. Bidder's Edge 2000)" as the threshold for action (emphasis added). To assess the potential for damage in *eBay*, the court applied a slippery slope argument (Kaplan 2000). "If the court were to hold otherwise, it would likely encourage other auction aggregators to crawl the eBay site, potentially to the point of denying effective access to

[9] The case was settled out of court prior to trial (Bloomberg News 2001).

eBay's customers (eBay v. Bidder's Edge 2000)." The court was therefore implicitly suggesting that *who* accesses the content can affect a finding of *possible, irreparable harm*. In the case of *eBay*, individual users might be desirable whereas integrators who are one, two, or many times removed from specific users are unwelcome.

3.2.4 Why

A single Web browser can choose to create a cache or not. Moreover, sophisticated browsers today can support separate caches for individual users or aggregate all user requests in a shared cache. Likewise, regardless of whether they process real-time queries (no cache) or create caches, integrators may act for individuals or aggregate users. Integrators that act on behalf of individual raise different policy concerns than those who aggregate users.

MP3.com is an on-line music listening and distribution service. "BeamIt," a new feature of MP3.com, was originally cast as an on-line "locker" service. In a conventional locker service, users upload music from their personal CD collection to a network host which then makes the recordings available to that single user from any network accessible computer (e.g. Myplay.com). In its purest form, the network host simply acts as a password-protected, network disk or an individual user cache. One variation on the theme might store music files in a system-wide database (shared cache) rather than in separate user directories to reduce redundancy. A service could even pass the efficiency on to users so that only the first listener to select "The Eagle's Greatest Hits" would have to upload the entire disc. Subsequent CD owners could, by verifying their ownership, gain access permissions to the respective files in the shared cache.

"BeamIt" extended the concept of the "locker" service to an extreme. MP3.com pre-loaded approximately 80,000 songs into a shared, on-line database (Hu 2000). By pre-loading the music, MP3.com not only economized on storage, but also spared users the cost of uploading an entire disk. UMG Recordings (Universal) filed suit claiming wholesale copyright violations by creating a commercial, digital music library from materials for which MP3.com lacked the rights (MP3.com 2000). "The only issue in the lawsuit is the propriety of MP3.com's having launched a commercial business with music it does not own and has not licensed (RIAA 2000)." MP3.com claimed that they were "simply facilitating a private consumer's storage of his or her privately purchased and privately used CDs (MP3.com 2000)." The court disagreed. "[F]actually, this purported justification was little more than a sham.... [Users] did not, in fact, store their own CDs or the sounds transmitted from their own CDs ... (MP3.com 2000)." MP3.com was found in willful violation of copyright and fined accordingly (Hu 2000; staff 2000).

The court's *MP3.com* decision in favor of the source providers focused on the issue of pre-loading. "[T]he difference between [BeamIt] and simple storage was critical to the anticipated commercial success of the new service (MP3.com 2000)." Of equal significance for integrators, however, was the finding that even were BeamIt to have required users to upload their own content, use of a shared, pre-loaded cache "does not meet a single one of the legal tests for 'fair use' (MP3.com 2000)." The court's finding against fair use is in keeping

with other decisions concerning third-party aggregation on behalf of consumers with regard to the use and redistribution of content. Consider the case of *Princeton University Press v. Michigan Document Services, Inc.* (Princeton v. MDS 1992; 1996).

MDS is a commercial copy shop in Ann Arbor, Michigan that produced academic course packs for classes taught at the University of Michigan, Ann Arbor. Professors would select a set of readings for a particular course and deliver the whole works, along with a course syllabus, to MDS. MDS would photocopy the assigned excerpts, and compile and bind them in an anthology. Students could buy the packs in lieu of purchasing the complete works, generally at a considerable discount. MDS did not make any attempt to contact the publisher's, pay fees, or otherwise receive permission from the copyright holders either prior to or following the creation and sale of a course reader. The decision to ignore the rights holders was a conscious protest against the decision in *Basic Books, Inc. v. Kinko's Graphics Corp.,* (Kinko's 1991) where Kinko's, a commercial graphics and printing shop, was found to have violated the copyright statute by creating course packs without permission.

To see the parallel to integration, we might think of MDS as an integrator. A single user, the professor, uploads content by submitting a query (course syllabus) and identifying data sources (originals). The professor downloads content by taking a single version of the course reader. A single use is permissible. Subsequently, students from the course visit MDS. Like a simple locker or storage service, students could individually submit a syllabus and course material to be copied, or like a shared cache, make use of the materials already in the MDS archive.

In finding against MDS, the court made several key points. First, the court endorsed a generous standard for evaluating an anti-competitive fair use when citing *Kinko's*, (Kinko's 1991 at 568), quoting *Sony*, (Sony v. Universal 1984 at 451). "[O]ne need only show that *if the challenged use 'should become widespread, it would adversely affect the potential market* for the copyrighted work (emphasis in original)(Princeton v. MDS 1996)." Second, the court concluded that even though an individual student or professor could have legitimately compiled the specific course readers in question, fair-use rights do not apply transitively to user agents. "[I]f the fairness of making copies depends on what the ultimate consumer does with the copies, it is hard to see how the manufacture of pirated editions of any copyrighted work of scholarship could ever be an unfair use (Princeton v. MDS 1996)." Third, the commercial nature of the third-party (MDS) weighted against them. "[T]he courts have ... properly rejected attempts by for-profit users to *stand in the shoes of their customers* making nonprofit or noncommercial uses (emphasis added, citing Patry, *Fair Use in Copyright Law*, (Princeton v. MDS 1996))."

Integrators deal with non-copyrightable data. By contrast, the decisions in *MP3.com* and *MDS* revolved around copyright. However, the decisions are still instructive for integration. The court establishes a slippery slope standard of harm in *MDS*. There are actions which, when performed on behalf of a single individual, are harmless. For example, copying a few pages or integrating a query. Those same actions, however, when multiplied over many users,

can result in actionable damages. Perhaps more significantly, acting on behalf of a single individual to avoid the slippery slope is no defense. What is permissible for a user does not necessarily extend to a third-party acting on behalf of that individual. From *MDS*, it is clear that commercial reuse is particularly suspect.

MP3.com and *MDS* together suggest that integrators who aggregate single queries over a number of users may face tight scrutiny. That scrutiny is likely heightened if the integrator is engaging in commercial reuse (i.e. *how* is the content used). Interestingly, where *MDS* suggests that acting on behalf of even a single-user may prove problematic, *MP3* suggests otherwise. Though not directly addressed, the court's language suggests that a pure storage service like a personal locker service might not have raised the same objections. Moreover, this contention is empirically born out by the absence of litigation against music locker services on the Web today.

3.2.5 How

From *MDS* we see that commercial reuse may weigh particularly heavily against an integrator. *MDS* therefore suggests the need for an additional set of policy instruments that govern *how* an integrator may use integrated content. Misappropriation or unfair competition finds its roots in the 1918 Supreme Court opinion *International News Service v. Associated Press* (INS v. AP 1918). INS and AP were competing news wire services. Barred from transmitting information on the Great War from Britain, INS reporters began to use AP stories published on the East Coast as a source for stories published on the West Coast, sometimes beating West Coast AP affiliates to press. AP sued on the grounds on the unfair competition. In light of the earlier discussion of *Feist*, it is worth noting that AP never sought relief on the grounds of copyright. INS stories were based upon historical facts gleaned from competing AP stories, and facts are not copyrightable (Spaulding 1998). In finding for AP, the Court identified three key points: investment of time and labor, market value of the product, and economic incentive to induce similar future work.

Two more recent expansions of misappropriation doctrine also concerned the news, this time with respect to sports. In finding for the State of Delaware, the court ruled that use of NFL scores in a lottery game was not in direct competition with the NFL and was therefore not actionable misappropriation (NFL v. Delaware 1977). Likewise, a Federal court found in favor of Motorola. Citing Justice Holmes in *INS*, the NBA claimed a "Hot News" misappropriation of broadcast rights by Motorola's SportsTrax service which sent NBA scores and game updates to pager owners (Djavaherian 1998). Absent competitive harm, the court concluded that there was no free-riding found no free-riding (NBA v. Motorola 1997).

3.2.6 When

Just as integrators may cache or not, the data with which they respond to users may be processed in real-time or delayed. Delay is introduced either by pre-fetching content from external sources so that the data used to answer a query is already old or by delaying a query request.

Real-time querying, by definition, calls for an integrator to pass user requests directly to underlying data sources. More precisely, governed by the number of queries fielded, the integrator would repeatedly query an external provider. This process, unregulated in the United States, might violate the EDD's restriction against "repeated and systematic extraction (Hunsucker 1997)."

While the European Database Directive may have some applicability to real-time queries, it was almost certainly crafted to directly address integrator pre-fetching. In order to maintain some measure of timeliness in the cache, refresh strategies require "repeated" access to external data sources, albeit typically on a longer time interval than necessitated by real-time querying. Populating a cache in anticipation of rather than in response to direct user needs would also likely require a comprehensiveness that would invoke the prohibition against "systematic" extraction.

The US currently has no legislation equivalent to the EDD, but responding to queries using delayed data can raise trademark concerns. We refer again to *eBay v. Bidder's Edge*. Of eBay's nine complaints, several, including false advertising and federal and state trademark dilution, stem from the observation that caching of data by a third party "can lead to outdated information about the current status of bids on [eBay], potentially harming eBay's reputation by confusing consumers (Krebs 2000)."

eBay's claim suggests that delay can lead to poor quality information and therefore disqualify reuse and redistribution. By contrast, in his concurring opinion on *INS*, Justice Holmes suggested the opposite. Calculated delays in reuse may sufficiently balance an initial producer's incentive to gather data against the public interest in widespread dissemination. Justice Holmes' standard, nearly a century old, is arguably more relevant today in a networked, wireless world of near instantaneous communication. Holmes suggested a time sensitive moratorium on INS publication long enough for AP to recoup AP's initial investment. Misappropriation, under both the majority and concurring opinions, is therefore a mechanism for regulating *how* integrators may reuse or redistribute content that is gathered from external sources.

Because of the need to maintain relatively timely data, *when* content is extracted may run afoul of EDD-like prohibitions against repeated, systematic extraction. Without such systematic extraction, cached data may become stale and consequently compromise the reputation of the underlying sources as in eBay's claim. Conversely, there may be some classes of data for which some temporary prohibition against extraction is required in order to induce gathering in the first place.[10]

[10] A delay in the right to redistribute in order to allow the initial gatherer to recoup costs was part of the origin for the term 'Hot News.' See references to *INS* in (NBA v. Motorola 1997).

3.3 Identifying the stakeholders

Our interest in the policy analysis is to assess the need for change. We began by laying out the problem space and reviewing the prevailing policy landscape. Here, we consider the stakeholders and their respective interests. This section begins by categorizing the different stakeholders. We then use the structure of the problem space as first defined in Chapter 1 to highlight particular stakeholder interests.

In the problem space of data integration, there are four categories of stakeholders to account for: data subjects, carriers, providers, and users. Data subjects are the identifiable individuals used to populate privacy-related data sets. Patient records and point-of-sale data are two such examples. As noted in Chapter One, because the data is privacy related, we omit these stakeholders as beyond the scope of this thesis.

Carriers are the individuals who facilitate the conveyance of data between different stakeholders. In some past instances, data services have been held responsible for the content that they transported though they had no knowledge of the content.[11] Because of a trend towards treating service providers in the manner of common carriers as well as the fact that carriers are a constituency not unique to the data integrator's problem space, we also consider carriers outside the scope of this thesis. In the following subsections, we consider the remaining stakeholder categories of providers and users and their corresponding interests.

3.3.1 Providers

We used the term "provider" in earlier Chapters without definition, trusting to context and the reader's intuition to make our meaning clear. Here, we attempt to draw clearer distinctions. Providers are those who make data available for consumption. Producers are one class of provider. Producers comprise the individuals and institutions who collect, compile, arrange, standardize, correct, index, update, and cross-reference data. A second class consists of providers who are also users. This is the class of integrators. Integrators reuse content and may also perform a number of value-adding functions. In addition to reuse, integrators might recompile, reformat, and harmonize. As a distinguishing characteristic, integrators gather data from other providers rather than from raw data sources. Note that an integrator may behave like a user to underlying data sources but may itself serve as a source for other value-adding providers.

Regardless of their status as producer or integrator, however, all providers tend to fall into some combination of three, distinct market models. We derive these market models from the data taxonomy originally described in Chapter One. In describing the different models, we identify costs and revenue streams consistent with Perritt (1996). Examples of each market model are provided. It should be noted that there are also providers who do not conform to

[11] There have been cases addressing whether bulletin board operators are responsible for copyrighted content trafficked on their sites (Langin and Howell 2000).

the market models. Government sponsored research and other public interest data production and provision, as noted in Chapter 1, is also beyond the scope of this research.

Perhaps the most intuitive market model is one where the data itself is the good being sold. In the world today, examples of such transactions abound. Both IRI and A.C. Nielson Company collect retail point-of-sale data daily, aggregate that data to produce region, state, or nation-wide marketing statistics. The Thomas Publishing Company produces the Register of American Manufacturers documenting more than 155,000 companies. Among other products, the McGraw-Hill Companies publishes the Standard & Poor/DRI's US Central Database (USCEN). More than 23,000 series of U.S. economic, financial and demographic statistics are included dating as far back as 1900 for conducting economic trend analysis.

While data transacted in this market model may exhibit some degree of time sensitivity and may include data from both government and private sources, the true differentiator seems to hinge on replicability. Data in this market is not necessarily sole-source. As evidenced by the market for retail point-of-sale data, competition may exist. However, the cost of reproducing data in this marketplace from original sources could prove prohibitive. Collections in this market exhibit large fixed costs and high barriers to entry. Much of the data exhibits at least some archival value (some non-zero half-life). In some cases, such as the USCEN, the historical data is often not replicable at all.

A second model in the electronic market for data involves the use of data to support transactions for other goods or services. Any number of financial services companies offer access to real-time financial figures and other sources of business intelligence analyses to induce customers to execute transactions[12]. In addition to publicizing sales of their own goods and services, participants in this second model might also engage in data integration to support comparison shopping of either their own product or those of another. FedEx, for example, is a $19 billion enterprise that includes the world's largest express transportation company and the largest surface expedited carrier.[13] As a feature of their services, users can estimate shipping costs based upon origin and destination address, weight, dimensions, pick-up date, and shipment modality. Following pick-up, users can track package delivery progress by entering per-package or per-shipment tracking codes. InterShipper, an information integration service that specializes in delivery and logistics, integrates rate estimates and tracking data from major shippers including FedEx, DHL, United Parcel Service, and Airborne Express to enable customers to compare options and prices[14]. Shipping, not data, is FedEx's core business. To the degree that FedEx considers itself competitive, integrators like Intershipper effectively provide FedEx with marketing and advertising.

In this model, some data is actually an artifact or a by-product of the transactions being executed. Regardless of whether the goods or services were marketed electronically or

[12] Consider firms such as Charles Schwab & Co., Inc. or Datek. Online Financial Services, LLC
[13] FedEx Corporation.
[14] Intershipper may be viewed at: www.intershipper.com.

otherwise, the data would likely have been collected anyway. Consider FedEx, which tracks packages irrespective of whether that data is made available to the customer over the Web. Cost of entry into this market model relative to a first mover, particularly with data that is a by-product of the core business, is therefore low. The data is independently replicable only to the degree that a second-comer actually entered the market for the good or service being transacted and then derived the data accordingly. Time sensitivity is largely a function of the corresponding transaction being conducted. The price of a financial instrument could change from second to second while that of a package shipment may not even change from day to day.

Third, a particularly visible electronic market for data is the one where, as in the second model, data is not the good or service being transacted. Rather, users themselves are the currency being transacted. Originally framed in terms of on-line advertising, data was a means for drawing users to view banner ads. Some financial information sites such as StockMaster.com began by using this model. Internet portals such as Yahoo, Infoseek, and Excite originated in this mode.

The model has evolved over time, however. AltaVista and other search services have discovered how to use the index and search results themselves as advertising. Sites can pay a fee to improve their scores in a user's search results. Many comparison shopping services have similar strategies. Retailers essentially pay for placement in a price list. A search list is then not unlike a click-through banner ad. The behavior, some consumer advocates argue, is not dissimilar to early versions of airline fare systems that were developed by the airlines themselves and defaulted to listing available flights in an order weighted to specific carriers.

In many ways, this third model is a hybrid of the first two. Although data constitutes the core tangible asset, the purpose of the service is to draw users to some external on-line or off-line transaction of goods or services. Banner ads evolved into click-through ads which in turn evolved into pay-for-placement search services.

3.3.2 Provider interests

Shapiro and Varian (Shapiro and Varian 1999) define the two key strategies for achieving success in information intensive industries as cost leadership and product differentiation. Product differentiation is sustained through the lesson, "know thy customer (Shapiro and Varian 1999)." Even though we earlier identified two distinct classes of providers, producers and integrators, it seems that all providers share the underlying fundamental goals: cost leadership and product differentiation.

While the overarching goals are similar, however, these goals have different implications depending upon the type of data in question. Therefore, we structure our review of provider interests in terms of market models. Within each market model, we consider the impact of the attribution-related problem space on the common strategies of cost and differentiation. In transitioning from provider interests to user interests, we return to consider differences between producers as providers and integrators as providers.

3.3.2.1 Data is the good or service

Where data is the good or service, perhaps the greatest provider concern is that of competitors who achieve cost leadership by free-riding on first-mover investments in database creation. The provider interest encompasses both *what* can be taken and *how* that content may be used.

Since 1948, Warren has annually compiled and published a factbook of cable system operators in the United States including name, address, number of subscribers, channels, provided, services offered, prices, and operator equipment. In 1989, Microdos began offering a competing, electronic product covering similar cities, using a similar set of data fields, and containing the same data. Warren Publishing filed suit against Microdos in (Warren 1997). That Microdos eventually prevailed on all counts motivates provider interests in *what* and *how*.

The principal difference between *Feist* and *Warren* highlights a second set of provider concerns where data is the good or service. In *Feist*, Feist was not competing directly with any existing service. Rural had no presence in the market for regional directories, on-line or otherwise. Feist, however, recognized a differentiable market segment within the broad market for directory information and sought to exploit it.

Product differentiation depends upon knowing your consumers and identifying opportunities for pricing, versioning, and other differentiation strategies (Shapiro and Varian 1999). As a consequence, knowing *who* is querying a particular data product becomes significant.

Moreover, common strategies for differentiating products include real-time versus delayed distribution of updates and variable pricing such as site licensing. Managing real-time versus delayed distribution requires control over *when* queries are executed or *when* the updates used to answer queries are processed. Delaying the re-use or re-distribution of data discriminates classes of users and prevents the delayed product from competing directly with the fresh data. Pricing policies like site-licensing require knowing whether a query represents a single individual or multiple users much as a musical recording might be purchased for personal use or public performance. In the attribution problem space, we categorize groups versus individuals as *why* a query is posed.

Providers are most interested in deterring cream-skimming, where second-comers identify high-margin users and then free-ride on someone else's initial data investment to develop a differentiated product that captures the high-end.

3.3.2.2 Data is a vehicle or advertising for the underlying good or service

There are times in which data, as part of the product, ceases to differentiate between competitors. Such was the case with real-time stock quotes when investment services first made their move into the on-line realm. Early on, services could offer real-time prices to separate serious investors from novices and price-sensitive experimenters. Over time, however, as first one and then other services began offering free real-time quotes, charging

for real-time quote became unsustainable. Ticker data became yet another component of the standard data set used by competitors to attract users.

Not all data in this market model evolves from data that originated as a good or service. On-line retailers routinely distribute databases of products, product descriptions, prices, and available inventory. In this market model, the distinguishing characteristic is that data serves to sell an underlying service.

Where the primary goal is short time-horizon sales of a specific good or service, knowledge of *who* and *why* are less significant. Whether the data is queried by a single individual, an individual acting on behalf of many others, or an on-line bot gathering data for a price comparison service, the principal concern is that the querying agent redirect sales back to the data provider who seeks to drive an underlying product offering.

Assuming that misrepresentation is not an issue, even if retailers are concerned with *how* pricing data is used (in particular price comparison services), early evidence in Internet marketing suggests that retailers can ill-afford not to participate in at least some manner of price aggregation. Little is lost from price-sensitive consumers who choose to buy elsewhere and more important is the exposure and effective advertising gained (see discussion below on the third market model for data).

Additionally, for providers seeking to gain cost leadership, *what* data is used may not prove significant. Data costs in this market model are often incurred as a part of providing the core service. Retailers would necessarily produce price lists and catalogs whether they were distributed on-line or not. As noted earlier, FedEx would gather shipment tracking data regardless of whether the data was shared with the public.

What does become significant in the context of the second market model, however, are the twin issues of *when* and *where*. First, *when* queries are processed and data is updated is significant when accuracy is required to drive the underlying good or service. Outdated data was at the heart of one of eBay's complaints against Bidder's Edge (eBay v. Bidder's Edge 2000).

Where data comes from to answer a query is relevant for two reasons. First, as noted earlier, use of a price-comparison service is not necessarily detrimental provided that users are directed to the correct source from which they may purchase the desired product. The link between *what* and *where* therefore requires accuracy. Second, and possibly in conflict with the need for accuracy where price aggregation services are concerned, is the question of *where* and how frequently *when* queries are processed. In particular, providers in this second market model do not want bots that are busy refreshing caches to exhaust system resources and introduce costly delays to end users seeking to purchase. Distinguishing between bots and human users lay at the heart of eBay's trespass claim against Bidder's Edge (eBay v. Bidder's Edge 2000).

The problem with limiting our analysis of providers in the second market model to short time-horizon sales of a specific product is the trade-off between one-time sales and building a long-term relationship with the customer. Providers in the second model who fail to identify *who* and *why* behind third-party integrators and comparison services lose the ability to cross-sell in the near term and the ability to further differentiate their products and consumers in the long term.

Identifying *who* and *why* motivated Ticketmaster's dispute with Microsoft on the issue of deep-linking. By bringing Sidewalk users directly to Ticketmaster's purchase pages, Ticketmaster lost more than the ability to manage the user's ticket buying experience and show banner ads. More significantly, Ticketmaster lost the ability to push related products and services and lost the knowledge of a specific user's browsing and searching patters for customizing future goods and services.

3.3.2.3 Users as the product

In this third market model, data serves as a way to deliver users, whether through banner ads, click-throughs, or search services. Because the entire market model is based upon delivering users, providers in this market model are particularly sensitive to the need to identify *who* and *why*.

Neither *what* is taken nor *how* the content is used is significant provided the data provider can deliver *who* and *why* in adequate numbers to their true customers. Indeed customization and differentiation in this market model is aimed at tailoring the environment to individual users based upon prior behavior.

The third market model is therefore like a hybrid of the first two. Customization in this market model is distinguished from differentiation of products (where data itself is the good or service) because the customer does not purchase data in this environment. Indeed data is given away in an effort to bring users to an advertiser or other service. However, the third model is distinguished from the second model because providers in the third model have no underlying service. Consequently, a provider in the third model lacks the same concerns about cross-selling.[15]

An important observation is that the third market model is heavily dominated by integrators rather than data producers. This suggests that among providers, producers are heavily driven by concerns over *what* and *how* while integrators, in their role as providers, are much more concerned with *who* and *why*. Such reasoning is borne out by support for the different legislative proposals reviewed in Chapter 2. Proponents of strong property rules to govern data reuse and redistribution are heavily dominated by producers while integrators and users tend to oppose strict regulations.

---

[15] Note that the provider does possess the crucial information about buyer behavior that Ticketmaster both lacks and desires. In this way, a third-model provider can become a first-model provider. They can sell comprehensive user search data. Because such markets raise significant privacy considerations, they are left beyond the scope of this thesis.

Tightly knit to the differences between producers-as-providers and integrators-as-providers is the reality that while integrators serve as information sources, they are also users. Consequently, we now turn to consider users and user interests.

3.3.3 Users and user interests

In Chapter 1, we suggested that users' interests in the attribution problem space were driven by quality and search. From the perspective of the different market models for data, however, we can define user interests in the attribution problem space in terms of switching costs in general and search costs in particular (Shapiro and Varian 1999).

Switching costs refer more generally to the costs of moving between different data providers. In some respects, what providers view as product differentiation is merely one way of locking-in consumers (imposing high switching costs) from a user perspective. Many value-adding services effectively raise switching costs. Custom data formats, data manipulation tools, interfaces, and related data sets are all ways in which providers can tailor products to specific users thereby making it more difficult to switch. In this context, quality metrics such as linking between *what* and *where* or providing users with meta information on *when* queries are processed and the timeliness of various sources are such differentiators.

Search costs are an attribution-specific switching cost. Certainly if a user wants to switch data providers, she must first identify a viable alternative. By documenting and directing users to links between *what* and *where*, attribution can ameliorate some of the switching costs. In particular, recall that attribution can help identify alternate derivations and equivalent sources for the same content.

Integrators as users, therefore, are interested in access to a large number of sources not only for their own sakes as users but in order to provide their users with as broad an array of underlying sources as possible, thereby possibly differentiating themselves from other competing providers. While integrators want to protect themselves from high switching costs, it is interesting to note that as providers, they have an interest in finding ways of not only differentiating themselves but also of locking-in their consumers and users.

3.4 The need for change

To assess the need for change, we now re-examine the policy landscape from the perspective of the different stakeholders. In doing so, we arrive at an emerging consensus; the status quo environment is inadequate to address the rapidly evolving attribution problem space. However, there remains widespread disagreement on the direction and degree of necessary change. We therefore consider, in turn, two arguments. First is the argument that producers have no protection and under the status quo are at the mercy of free-riding integrators. Second is the argument that producers have all the advantages, and that status quo policies are biased against integration and other consumer-oriented data services provision.

3.4.1 Free riding integrators have the advantage

Producers, recall, have a particular interest in limiting *what* and *how*. As a consequence, they are particularly troubled by decisions that either explicitly constrain or implicitly weaken a producer's legal right to govern *what* and *how*.

Of particular concern is the Court's decision in *Fiest* and the derivative cases that followed that, by default, place comprehensive, logically (intuitively) organized collections of facts into the public domain.

Contracts, as in the case of *ProCD*, offer no safe haven (ProCD v. Zeidenberg 1996). First is the question of whether contracts can pre-empt the Constitutionally rooted copyright basis for the Court's decision in *Feist* (Elkin-Koren 1997; Ginsburg 1990; 1992). Second is the observation that contracts are only enforceable against parties to the contract. Even though a specific individual user may be found guilty of violating a contract against commercial reuse and redistribution, no third party is equally liable. Were a third party, not under contract, to obtain a copy of the data, he would have no contractual obligation to refrain from commercial reuse.

Finally, to the question of *how*, even though *INS* suggested that direct competition was prohibited, subsequent cases that derive from *INS* have proven quite inconclusive. In an interesting foil to the *NBA* and *NFL* cases, a third sports case found in favor the initial broadcaster. Transradio Press Service (TPS) used spotters and the ringside fight announcers as sources to broadcast boxing matches sponsored by Twentieth Century Sporting Club. TPS was competing directly with NBC, who had an exclusive contract with Twentieth Century for radio broadcasts. Because TPS was using the NBC broadcast as a partial source, the court found unlawful misappropriation by TPS (*Twentieth Century Sporting Club, Inc. v. Transradio Press Service*, (Transradio Press Service 1937)). The variations in outcome emphasize the fact that while the Federal courts may pass judgment on misappropriation cases, there are no Federal laws regarding misappropriation. As a consequence, though *INS* was a Supreme Court decision, the Federal courts have long since been forced to rely upon (wildly inconsistent) State laws (Spaulding 1998).

3.4.2 Producers have the advantage

By contrast, integrators and other value-adding innovators point to a host of undecided cases or cases settled out of Court and argue that the specter of strong property rights in line with the European Database Directive would stifle future innovation in the development of data products and services.

First, integrators point to the appellate court decision in *ProCD* to raise the potential for shrink-wrap licenses and contracts to preclude wrapping-based data aggregation as a user-centric service. The threat of prohibition is only magnified by trespass claims, untested though introduced in *eBay*.

Caching as a strategy both for performance and for aggregating data over a number of users was challenged both in *eBay* and in *MP3*. Absent a decision, integrators are perhaps faced with the precedent from *Princeton v. MDS* which, although it concerned copyrightable materials, established two points that could carry over into the realm of data reuse. First, *MDS* established that a commercial service could not necessarily serve as an agent for or "stand in the shoes" of another. Second, *eBay* suggested and *MDS* established a slippery slope argument with respect to commercial reuse. The principle states that although a single use might not prove abusive, because the same act multiplied over hundreds if not thousands of users could stifle initial investment incentives, prohibiting the single use is justified.

Finally, although *INS* suggested that only use in direct competition with an initial provider is prohibited, both producers and integrators at least agree on the irresolution offered by misappropriation.

Though the different stakeholders disagree on the nature of the necessary change, they at least agree on the need for some measure of intervention. If for no other reason, U.S. inaction has ceded the field to the European Database Directive (EDD). Integrators find the EDD too restrictive and would like a counter-proposal. Producers favor the EDD and point to the reciprocity clause requiring parallel US legislation if domestic data producers are to receive equal protection in European venues. With the need for change in mind, we now turn to Chapter 8, an exercise in policy formulation, to address the attribution problem space.

4 Policy formulation

The Policy Analysis of Chapter 7 leads us to conclude that some form of Federal intervention into the arena of databases and property rights is inevitable. Building from this assumption, we conclude that a Federal misappropriation statute for database production best serves the Congressional mandate to "promote the progress of science and the useful arts (U.S. Constitution, Art. 1, Sec. 8)."

Our basic contention is that databases are a unique form of intellectual property. The disaggregation of content and presentation (Bray, Paoli, and Sperberg-McQueen 1997; Walsh 1997) made possible by modern information technologies enables the separation of fact from "selection and arrangement" in a way that could not have been foreseen when the framers crafted the Constitution (Feist v. Rural 1991). The Court had it "right" in *Feist*. Attempts to claim a property right in data are mired in the print-and-paper-based past.

The decision in *Feist* prescribed copyright protection to selection and arrangement. Some view the Court's refusal to apply similar protection to facts as a denial of any protection for the 'sweat work' involved in gathering and collecting data (Duncan 1999; Horbaczewski 1999). Instead, perhaps the Court was merely calling for the Congress to perform its duty and legislate a misappropriation right rather than asking the Court to "create policy," echoing the admonition made by Justice Brandeis in his dissent to *INS v. AP* nearly a century before (INS v. AP 1918).[16]

In Section 1, we ask the question, "why do we protect intellectual property?" The policy that we propose and the mechanisms that we select depend, in part, on what we aim to achieve through protection. Therefore, we begin by asking what goals Congress should seek to fulfill. We conclude that ideal policy proposals to address the attribution problem space are those which best promote innovation.

In section 2, we then ask, "what are we trying to protect?" Where is there room for innovation in databases? We separate a database into two distinct elements, the product of a creative process in selection and arrangement, and the product of a laborious process in gathering. In part, our purpose is to dispel the apparent misconception that "some compilations, particularly computerized databases, may lack any 'arrangement,' for they are designed to permit the user to impose her own search criteria on the mass of information (Ginsburg 1992 at 346)."[17] In the end, we conclude that one element of the database is protected under copyright while the second element is left unprotected. The remainder of the Chapter then considers protection for the unprotected products of gathering data.

---

[16] In his dissent, Brandeis wrote that "Courts are ill-equipped to make the investigations which should precede a determination of the limitations which should be set upon any property right in news or of the circumstances under which news gathered by a private agency should be deemed affected with a public interest. Courts would be powerless to prescribe the detailed regulations essential to full enjoyment of the rights conferred or to introduce the machinery required for enforcement of such regulations (INS v. AP 1918)."

[17] See also (Patterson 1992 at 395).

Having clarified what we are trying to protect, in Section 3 we ask, "how do we protect it?" We examine two different economic frameworks that have been applied to the study and management of intellectual property. The first framework is the standard Prisoner's Dilemma (Gibbons 1992) and the second is the legal entitlements framework first crafted by Calabresi and Melamed (1972). Each framework serves as a theoretical benchmark for evaluating both the need for change and the viability of our policy formulation.

In Section 4 of this chapter we ask, "what is so special about data?" Combining observations from our Chapters developing a formal model of attribution and from our Policy Analysis, we identify some significant differences between databases and other forms of intellectual property that may challenge the correctness of applying general conclusions about the management of intellectual property rights to our specific question: balancing value-added innovations in the market for databases (i.e. data integration) with the producer's incentive to create databases in the first place.

Section 5 of this chapter documents our proposal for a Federal misappropriation statute as a legislative strategy for addressing the balance between re-use and re-distribution versus production. A three-part operational definition is provided that also serves as a test to justify a plaintiff's claim of misappropriation. Potential remedies are also considered.

Section 6 contains an evaluation of our policy proposal with respect to the two theoretical frameworks laid out at the beginning of this Chapter. Common criticisms of the misappropriation doctrine in general and elements of our proposal in particular are raised in Section 7. In addition to theoretical arguments, we address pragmatic considerations about implementation.

4.1 Why do we protect intellectual property?

There are many reasons that have been proposed for why we protect intellectual property. There are arguments that a property right in their work is a natural right inherent to creators or that it will better promote the free expression of ideas (Merges et al. 1997). There are economic arguments that granting rights will induce authors and inventors to write and create or that doing so will stimulate trade (Posner 1992). The perspective that we take in this Chapter is that the purpose of protecting intellectual property is innovation: the stimulation of new works.[18]

Our motive for selecting innovation as the motive for intellectual property protection stems from our interest in studying legislative remedies to the challenges presented by the attribution problem space. Legislative action is justified by the Constitutional mandate

[18] This is not to suggest that other perspectives are incorrect or that there is nothing to be gained from adopting an alternative perspective. Indeed there are philosophical arguments on the mutually reinforcing or contradictory natures of these different goals. A different assumption could very well lead to a different conclusion, however, and identifying and reconciling those perspectives is a different study.

defined in Article 1 Section 8 to "promote the progress of science and the useful arts (U.S. Constitution)."

Innovation from an initial creator is quite straightforward. Examples of initial creation are the inventor of a new product, the author of a new story, or perhaps the compiler of data that has never before been systematically ordered. This is typically understood as an argument to protect and/ or grant rights in order to promote original creation.

However, progress and innovation do not end with creation. We can think of 'new creations' and incremental improvements. Invention begets invention. Creation does not take place in a vacuum (Merges et al. 1997). All "new" works in one sense or another builds upon prior progress (NRC 1997a). Intellectual property protection, to borrow the application of the term from Ginsburg, is a sauce that covers the follow-on goose as well as the initial creating gander (Ginsburg 1990). Protection provides the same incentive to creators that build from the existing pool of knowledge and creation.

Incremental improvement offers a second level of innovation. "One person invests labor and money to create a product, such as a food processor that people will buy. Others may imitate him and take advantage of the new market by selling their own food processors. Their machines may incorporate their own ideas about how such machines should be made. As a result, the quality of the machines may rise and their price may fall.... [T]he public as a whole may be better off (Baird 1983 at 415)."

It is, in fact, the essence of intellectual property protection to protect and promote not only original creation but also follow-on works (O'Connor in Feist v. Rural 1991). Innovation, then, is what Congress is charged to promote. The question facing legislators, then, is where does the innovation in databases lie?

4.2 What are we trying to protect

For policy purposes, a database is defined as discrete facts, data, or other intangible materials collected or organized in a systematic way in one place or through one source so that users may access them (H.R. 354 1999; H.R. 1858 1999). Our contention is that a database entails both a creative design component and a labor-intensive sweat component. In print-based media, the two types of work are inextricably intertwined in the final product. However, we argue that modern information technologies have enabled the disaggregation of creative work and sweat work. Creative works ("selections and arrangements") are protected by copyright. What remains is the question of protecting the sweat work (the disaggregated "data").[19]

4.2.1 Database design: selections and arrangements

We begin by arguing that database design is a distinct process. This process occurs independent of the medium in which the product is ultimately rendered (e.g. in print versus

[19] References to "selection and creation" and "data" are to the Court's ruling in *Feist* (Feist v. Rural 1991).

electronic form). The practice of database management systems separates database design into three modeling tasks: conceptual models, logical models, and physical models (Rob and Coronel 1997). Loosely framed, the conceptual model defines scale and scope (Ramakrishnan and Gehrke 2000; Rob and Coronel 1997). By scope we refer to the elements, attributes of those elements, and relationships between those elements. In the database represented in Table 3.1, we captured information about lodging, transportation, and tourist attractions. Hotels have attributes like name, address (geographic location), and room rate. Tourist attractions have names and are located in specific regions. By scale we refer to the extent or quantity of data in the system.[20] The database of Table 3.1 includes data in and around the city of Tokyo, Japan.

Some other tourist database might choose a different scope. For example, restaurants instead of or in addition to tourist attractions; amenities like hostel meals or hotel health clubs as an additional attribute of lodging establishments. Tourist databases might also differentiate themselves conceptually on scale. There are some guides for cities like Tokyo and others for the entire country of Japan. Some guides focus on students and other low-budget travelers (Let's Go 1993; Planet 2001) while some target businesspersons and the well-heeled.

A logical model defines the organization or framework for ordering the data elements, attributes, and relationships selected in the conceptual model.[21] As with conceptual models, this organization has two dimensions. First, the collection has a fixed arrangement or "schema." Certain information, like rooms and prices, are in one table. A hotel's geographic information is in a different table. Geographic information on tourist sites are in yet a third table. Second, each table itself has a distinct ordering.[22] Name is followed by room-type, which is followed by price. The implicit interpretation is that, for any given row, the price corresponds to the room-type at the associated hotel-name. Because we read from left to right, it makes sense that names are on the left rather than prices.

Some other collections of tourist information can and do arrange information differently. All low-budget items could be in one category and all high-budget categories could be in a second. Alternatively, information could be principally ordered geographically rather than by separating hotels and tourist attractions. A single table could list regions and all of the attractions, lodging, and transportation within that region.[23]

A physical model describes how the data is ultimately rendered. In particular, the logical model is translated into some literal format on paper or disk that is optimized for a particular

---

[20] In the relational context, this is formally defined as the finite subset of the Cartesian product of finite or countably infinite domains that comprise a relation. See Ullman (1988) and the text in Chapter 5 of this thesis.
[21] In the relational context, this is formally defined as the schema. See Ullman (1988) and the text in Chapter 4 of this thesis. In the industry, this component is referred to as the process of schema or database design. See (Ullman and Widom 1997) and the following text on data modeling.
[22] Formally, of course, order does not matter. The set of lists notation where order matters is equivalent to the set of mappings (Maier 1983; Ullman 1988).
[23] The reader may object at this point that there is no reason a single guide could not provide all of these orderings. We address this issue in the text below.

kind of a query. In the same way that the logical structure enables or precludes certain types of queries, the corresponding physical model can affect the speed or efficiency with which certain types of queries and operations execute. Consider, for example, the physical format of the Yellow Pages. It facilitates search by subject area and only secondarily by alphabetical ordering of company name. Searching by region, as supported in our hypothetical electronic travel guide, is not supported by the Yellow Pages' physical data model.

Commercial database software largely makes the issue of designing physical storage a moot point. Most commercial software vendors use some variant on a balanced tree (Ramakrishnan and Gehrke 2000). The important point is that different logical structures are translated as different physical trees. As an aside, we observe that it would be wrong to conclude that there is no creativity in physical modeling. Indeed the competitive environment between Oracle, Microsoft, Sybase, Filemaker, and other large and small scale database software vendors, who all build on the same logical framework[24] suggests that there is ample room for creativity at the physical level.[25]

Modern database design entails conceptual, logical, and physical modeling. The process of design occurs wholly independent of the process of gathering data and placing it into the framework. By separating the selection in conceptual modeling and the arrangement in logical modeling from the process of gathering data, the Court's ruling in *Feist* acknowledged the clear distinction (Feist v. Rural 1991).

4.2.2 Creativity and sweat in database creation

Having argued for a distinction between the process of database design and the process of gathering data, we next consider the balance of intellectual creativity and brute sweat in the two. First, we argue that the creativity in database design is non-trivial.[26] Good design depends upon a set of mathematical normalization rules and upon expert knowledge of what prospective users intend to query (Ramakrishnan and Gehrke 2000; Rob and Coronel 1997). Conversely, poor design can contribute to unnecessary repetition, data inconsistencies, and may prevent the ability to pose certain queries altogether. Good design is the heart of the intellectual creativity in database creation.

Consider again the travel database from the Introduction. The reader will note that in documenting price, we did not identify currency. We implicitly assumed that Japanese prices would be listed in Japanese Yen. However, this is only a thesis example. Even a cursory review of on-line and print guides would quickly reveal that not all prices are reported in Japanese Yen. Foreign guides for Tokyo might report in local currencies (e.g. U.S. Dollars, British Pounds), and international chains might always report in a single currency (e.g. U.S. Dollars) (hotelguide.com 2001; Japan Youth Hostels 2001).

---

[24] Most commercial database vendors implement some version of the relational data model at the logical level.
[25] While most vendors use some variant of a b-tree, they do compete in areas such as query optimization, query processing, data integrity checking, transaction processing, etc.
[26] The tongue-in-cheek reasoning argues that if database design is trivial, why do consultants get paid so much money to design them.

More crucially, the reader might note that rather than document hotel and tourist attraction addresses, the database of Table 3.1 identifies regions. Moreover, hotel regions are in a separate table rather than stored with other hotel attributes like room size and rate. In this stylized example, had we stored addresses rather than regions, we would have been unable to process queries like that of Q2 in Chapter 3 seeking hotels around the Imperial Palace. Had we stored hotel geographic information in the same table, we would have unnecessarily repeated the same address or region for every different room and rate.

The convention appears to accept that design is trivial. "[C]omputerized databases, may lack any 'arrangement,' for they are designed to permit the user to impose her own search criteria on the mass of information (Ginsburg 1992 at 346)."[27] However, we argue that there are at least three reasons that good design, the intellectual creativity in database creation, is non-trivial. First the relationship between data like street address and region or hotel geographic information and hotel room rates, in the example above, is captured in what are formally called functional dependencies. Functional dependencies are not discovered by exhaustively searching through large sample sets of data (Ramakrishnan and Gehrke 2000; Ullman 1988).[28] Identifying functional dependencies for database design requires domain expertise.

Second, good database design requires understanding what prospective users are interested in. Novel applications of the same data build from different logical and conceptual models of the same set of facts. Consider, for example, epidemiological studies of disease that mine longitudinal patient records (NRC 1997b). Doctors use (and consequently model) data in a patient-centric way covering all symptoms in reverse-chronological order. An epidemiologist studying a specific form of cancer might focus on only a subset of the data (selection), ordered by symptom or diagnostic test (arrangement), which is translated into a different physical format.[29] Consider also the difference between the yellow pages and the white pages business directory listings. Much of the underlying data is the same. Indeed the user populations are even the same. However, the use model for each directory is quite different. Finally, as an extreme, even two different database designers, given the *same* set of users and the same set of data, could arrive at equally viable but distinctly different underlying logical models.[30]

A third reason that database design requires creativity and is non-trivial is that functional dependencies and schema design can be difficult to decipher from the data alone. Looking only at the results of a query like Q2 of Chapter 3 does not easily suggest a good logical

---

[27] See also (Patterson 1992 at 395).

[28] Functional dependencies are formally a property of the underlying data domains. Exhaustive searching of data sets can reveal contradictions, but no data sample can prove that a functional dependency holds.

[29] The problem of unmaterialized views, akin to constructing a logical model without a corresponding physical model, is captured in the context of integration systems that do real-time querying of third-party sources and dynamically generate results. Our medical example is provided merely as an example of how different needs drive different logical models. Issues related to real-time queries are discussed below.

[30] The truth of this is repeatedly demonstrated in problem sets for classes on database management systems. Students, beginning with the same parameters, can arrive at imaginative solutions that differ significantly.

design. Even a standard white-pages directory listing, which seems obvious, may embed alternative orderings. There are first name orderings and both geographically and alphabetically ordered reverse-listings based upon address. Carefully defined user interfaces hide what users do and do not see and limit what users can and cannot query. The user, in asking Q2, does not know that the database of Table 3.1 uses region classifications for identifying proximity. While attribution might reveal the use of regions, it need not identify the schema design that separates hotel regions from other hotel characteristics.

Contrasting the heavily creative process of database design is the process of data gathering and manipulation. This process entails not only literally collecting data but also ordering that data in a consistent form and then verifying and updating content (McDermott 1999; Perritt 1996).

Collecting data invokes visions of U.S. Census takers going door to door, biologists in a lab counting cells beneath a microscope, surveyors measuring property boundaries, or grocery clerks recording items on the shelves to reconcile inventories. There is a distinct element of labor. Not even data collection is untouched by information technologies, however; without meaning to digress into a study of data collection, we observe that there is a continuum in data collection practices that range from the heavily labor intensive to the highly automated. Government-on-the-Web may reduce the pavement pounding required to gather census data, GPS can match surveyors on the ground, bar-code readers help reduce the costs of inventory management. The principle contention, revisited below, is that even the seemingly mundane task of gathering data is not without room for innovation.

In addition, the same innovations that impact the process of collecting of data may also apply to verifying and updating content. On a first order, verifying may involve revisiting original sources to ensure that data was captured correctly. For example, digital tools are a boon to the law review editor or legal clerk asked to verify citations. More generally, the process involves using (alternative) sources to confirm or contradict recorded data, recognizing that depending upon the facts in question, data changes over time. The same tools for gathering may therefore be applied for (re)confirming or updating content subsequently.

As an aside, it is worth noting that data collection does not occur in a "selection-arrangement vacuum." Distinguishing the process of creative selection and arrangement does not mean that the gathering process lacks any organization. First, any collection inheres selection by virtue of what is not collected. Indeed the initial database producer likely has prospective users and uses in mind, and it this set of needs that drives her selection. Second, by design, systematic data collection implies a certain structure, albeit one that is "practically inevitable" and not "remotely creative (O'Connor in Feist v. Rural 1991 at 1296-7)." However, the point in data collection is not to be original but to be rigorous. It is this rigorous consistency that

allows producers to treat a raw data collection as an input to the second process of creative selecting and arranging.[31]

Our contention in drawing a distinction between the two processes is not intended to suggest that the latter does not entail creativity. Indeed it is the very observation that there is a place for creativity in data collection that informs our subsequent policy proposal.[32] However, for the purposes of distinguishing the two processes, it seems uncontested by both proponents and opponents of database rights legislation that data gathering is heavily balanced towards laborious sweat.[33]

4.2.3 Databases in the print media

Though we argue that the two processes are distinct, it is also our contention that, in the print-on-paper world, the *product* of the data gathering process is inextricably intertwined with the *product* of selecting and arranging. A producer cannot render data without committing to and revealing a particular selection and arrangement. Likewise, one cannot use an alternative selection or arrangement without physically rendering the alternate arrangement as a separate print product.

Consider again the White Pages as a published database of names, addresses, and phone numbers. As noted earlier, it is possible to conceive of a number of alternatives to the conventional, last name-first name alphabetized ordering of listings. The process of selecting and arranging may reveal multiple products (conceivable orderings). However, the product of the data gathering process, presenting the data itself, is necessarily tied to and cannot be transferred without embedding *one* particular selection or arrangement.

This is not to suggest that the print media is incapable of representing alternative selections and arrangements. Local restaurant guides, for example, often present multiple arrangements of their selection. There are alphabetical listings by restaurant name, by cuisine, by geographic location, or perhaps even by special services (Brown 2000; Kravitz 2001). However, it is our contention that first, each index constitutes a distinct collection.[34] Second, alternative arrangements quickly become too exhaustive to print in a single publication for any database of significant size.[35] While a restaurant guide might provide an index that constitutes a different arrangement of the same selection, more complex collections such as

---

[31] Automated parsing of data, where for semistructured data querying or for loading into a relational structure, assumes a certain perennity to the data (Lee and Bressan 1997).

[32] See text below on why the market for data is different from other forms of intellectual property.

[33] Basically all parties, whether proponents or opponents of rights legislation, characterize the gathering of data as laborious sweat work (Corlin 1998; Hammack 1998; Tyson and Sherry 1997; Winokur 1999).

[34] To be sure, separate indexes are interrelated, perhaps by page number or restaurant name. The Zagat Survey (Brown 2000), for example, lists restaurants and associated attributes by alphabetical ordering of restaurant name and then presents alternative indices (cuisine, geographic location) by listing only restaurant name. (Kravitz 2001) provides restaurant names and page numbers. In relational database terms, each alternate index is a separate relation where restaurant name or page number constitutes a foreign key.

[35] Reverse telephone directories that list telephone numbers by geographic address are printed as separate documents.

travel guides often provide only abbreviated indexes that represent a more limited selection in an alternative arrangement.[36]

More significantly, not only is it costly to render different arrangements in print form, but also recall that each arrangement embeds a particular set of assumptions about user interests and search criteria.[37] Using a traditional White Pages directory to search by first name or a geographically ordered travelguide to search for a specific restaurant based upon the restaurant's name is largely an exercise in futility.

Consequently, in protecting a printed collection, it is not strictly necessary to distinguish between the data and the selection/arrangement as the object of protection. Protecting printed data inherently extends to its selection and arrangement. One cannot extract and use one without the other. However, the print media does not necessarily equate data and presentation.[38] Past technical limitations merely clouded the issue that eventually came to a head in *Feist* (Feist v. Rural 1991).

4.2.4 Electronic databases: disaggregation and appropriability

Modern information technologies disaggregate the product of gathering from the product of selecting and arranging.[39] There are many ways to arrange the same selection of data[40] and we might combine selected subsets from different collections to create a new whole.[41]

In the print media, users are prevented from using the same set of data in an unspecified way. For example, one cannot (easily) use the White Pages for a first-name lookup. However, database technologies, break down the apparent barrier to protecting alternative selections and arrangements posed by the print media. Different user populations can view a single set of gathered data through the lenses of different selections and arrangements. Likewise, the same schema definition or arrangement can be applied to different data sets.[42]

---

[36] Travelguides, which detail not only restaurants but also locations of interest, lodging, transportation, etc. typically do not offer a rich array of alternative indexes. When included, the available index often mixes a limited selection of locations of interest, lodging, etc. mixed together in a single, alphabetically ordered listing. See (Let's Go 1993; Planet 2001; Taylor et al. 1997)

[37] See text on Creativity and sweat in database creation (Section 8.2.2)

[38] Modern database technologies provide the languge for specifying selections and arrangements without sharing the data. Consider our earlier description of possible alternative arrangements of restaurant listings or travel information. More formally, we might refer to alternative arrangements as intensional databases or view definitions. See (Ullman 1988).

[39] To be sure, the data in a database conforms to a particular conceptual, logical, and physical model stored on a computer harddisk. Likewise, the user of a collection assumes a particular selection and arrangement to query over. The issue is flexibility in use.

[40] Some rearrangements are easily constructed as intensional databases of the original. Others are harder to define and may require additional data gathering because, for example, the original design might have omitted a necessary foreign key. See (Levy, Rajaraman, and Ordille 1996; Rob and Coronel 1997).

[41] See Chapter 1 of this thesis and (Wiederhold 1992).

[42] Data integration, more broadly, is exactly the process of redefining different data sets in terms of the same schema. See: (Levy 2000).

Disaggregation is enabled through the power of abstract data definition and manipulation languages (Abiteboul, Hull, and Vianu 1995; Maier 1983; Ullman 1988).[43] In relational database systems, a single SQL (Structured Query Language) instruction both selects and restructures (arranges).[44] [45] The threat is therefore not that users can create and distribute error-free copies with a point-and-click gesture. The true threat is that posed by the costless selection and restructuring capabilities of modern query languages.

But relational database systems have been around since the 1970's. Why did the apparent threat to commercial databases not appear sooner? The answer is the World Wide Web. What was hidden behind the arcane syntax of SQL was exposed via Web browsers on millions of desktops around the world.[46] The indecipherable foreign langauge of the relational data model is being supplanted by the semistructured data model of XML (Extensible Markup language), made accessible to users through their Web induced familiarity with HTML (Hypertext Markup Language).[47] Indeed one of the motivating themes behind XML is the explicit separation of content (data) and presentation (selection and arrangement) (Walsh 1997). XSL (Extensible Stylesheet Language) and emerging XML query languages like XQuery are to XML what SQL is to the relational data model (Chamberlin et al. 2001a; Chamberlin et al. 2001b; Fernandez and Marsh 2001; Fernandez and Robie 2001; Lenz 2001). XML query languages enable users to select and restructure data encoded in XML (deBakker and Widarto 2001; Katz 2001; Lenz 2001). Current innovations that extend beyond even XML, such as Microsoft's .NET and the Web Services initiatives, merely highlight the role that such schemas and interfaces will play.

The vision is that future Web content will be encoded in XML. Different users may then access the same physical data set XSL or XQuery instructions will then allow individual users to render customized selections and arrangements of the same physical data set through their desktop Web browsers (Abiteboul, Buneman, and Suciu 2000; Lenz 2001). Likewise, heterogeneous content will be integrated from physically distributed data sets using a similar set of instructions (Chawathe et al. 1994; Levy 2000; Wiederhold 1992). Modern information technologies emphasize the distinction between products of data gathering and products of selecting and arranging.

---

[43] Database Mangement Systems, Ullman, Maier for definitions of database definition and database manipulation languages.

[44] SQL stands for Structured Query Language. In the standard SELECT FROM WHERE syntax of SQL92, a user can define a creative selection or subset of data from an existing database by crafting an appropriate set of constraints. The FROM clause indicates which tables to extract data from. The WHERE clause indicates which data to take and which data to ignore. The SELECT clause structures the output into a new arrangement or table (Ramakrishnan and Gehrke 2000).

[45] It should be noted that there are some logical restructurings that are not supported by a query expression.

[46] Early standards like CGI (Guelich, Gundavaram, and Birznieks 2000) enabled users to integrate databases with the ubiquitous World Wide Web and has helped drive the evolution of new standards for representing and presenting data (Abiteboul, Buneman, and Suciu 2000).

[47] HTML is the Hypertext Markup Language, the current standard for rendering content within a Web browser. XML is the Extensible Markup Languages. The impetus behind XML was largely to replace and correct perceived limitations of HTML (Bray, Paoli, and Sperberg-McQueen 1997; Walsh 1997).

It is worth noting that proponents of strong protection for databases implicitly acknowledge the distinction between the two products and processes. "Competing firms rarely supply the 'same' database. Rather they compete on a range of fronts: selection of data; convenience; search engine; ease of use; and price (Tyson and Sherry 1997 at note 31)." We have hopefully demonstrated that variables like selection, search/query engine, convenience-ease of use (i.e. tailoring database schema design to the needs of particular users) are the products of a distinct process.

4.2.5 Database protection: selecting and arranging vs. gathering

As observed earlier, intellectual property protection is about balancing pressures for production against pressures to innovate. The Policy Analysis of Chapter 7 described strong arguments both in favor of and against the need for legislative intervention to restore this balance with respect to databases. Some argue that the status quo is sufficient and others press for action. Our conclusion, that a database is actually the product of two distinct processes, suggests a new interpretation of the analysis.

We observe that arguments pro and con largely aim at two different products. "Companies and interest groups have chosen sides on the issue depending on whether they primarily collect data that is put on the Internet (the stock exchanges, real estate brokers, Lexis-Nexis, eBay, the A.M.A.) or use the data compiled by someone else (the Chamber of Commerce, Consumers Union, Yahoo, Schwab, research librarians) (Rosenbaum 2000)." Those promoting the status quo focus on innovation in the second process, that of gathering. Arguments for change address incentives in the first process, that of selecting and arranging. The two are not inconsistent.

Stakeholders in the process of selecting and arranging argue in favor of the status quo (Bloomberg 1996). In Chapter 7, we identify several markets and industries that are built on the ability to (re)arrange and re(use) data in novel ways. These intermediaries are themselves database creators and suppliers, facing the threat of re-use and re-distribution (Ginsburg 1990). Yet their role as customers and users provides sensitivity to issues of access. Existing measures are, for these user/producers, sufficient for protecting their creative investment. Moreover, independent of commercial value, these stakeholders argue for the need to preserve basic factual data as a public resource.

Conversely, the analysis from Chapter 7 suggests that parties promoting protection focus on their investments in gathering. The fruits of investments in gathering are vulnerable to "the ease and speed with which a database can be copied and disseminated in the digital age (Monster.com 2000)." Of those who engage in data collection, there are typically three perspectives on the creativity and selection and presentation. First, there are those who argue that selection and arrangement is not a relevant concept in the electronic environment. "But to treat these acts as authorship for computer databases is a fiction. Within the database there is no coordination or arrangement (Patterson 1992 at 395)." Second, are those who imply that, while a distinct process, selecting and arranging often embodies little creativity and is

virtually costless (Ginsburg 1992 at 345). We have hopefully addressed these first two perspectives earlier in the text on creativity and sweat in database creation.

A third perspective is advanced by some database producers such as the National Association of Realtors (NAR). The NAR implies that the selection and arrangement of , their Multiple Listing Services (MLS) embeds information and expertise that can only be interpreted by experienced users (in this case, realtors belonging to the NAR). Pirates who redistribute MLS content without the ability to interpret the knowledge embedded in the selection and arrangement therefore place consumers at risk (Cronk 2000; McDermott 1999). Consequently, content protection is justified. However, the Court in *Feist* was quite clear that "even a minimum standard of creativity" in selection and arrangement would invoke copyright protection (Feist v. Rural 1991). That a selection and arrangement embodies expertise is not difficult to imagine. This is precisely our argument: that there is value in the process of selection and presentation. It is difficult to imagine how any selection and arrangement that embeds such knowledge and expertise would fail to qualify for copyright protection.

We are therefore left with two positions that largely pit primary producers with intermediaries in the market for data (re)use and (re)distribution. However, we argue that the two positions, which reflect the distinct processes in gathering versus selection and arrangement, are not inconsistent. In the past, when the products of the two processes were intertwined, protecting one implicitly protected both. Today, the challenge is to address the products of each process separately.

Some stakeholders, users and intermediaries who re-use and re-distribute data, are largely concerned with the products of creation and selection. For the purposes of protecting creation and selection, at least some feel that status quo protection is adequate (Bloomberg 1996; Perritt 1996; Shapiro and Varian 1999). "Bloomberg finds the existing combination of copyright law, contractual limitations, administrative practices and technological security to be adequate at present to protect its commercial interests (Bloomberg 1996)." In any event, we argue that creation and selection is a process distinct from gathering. Further consideration of appropriate protection for the creativity in creation and selection is left for future work.

However, the policy analysis in Chapter 7 also raised a host of objections to existing protections. These objections concern a second distinct process, that of data gathering. Lacking the Constitutional authority underlying copyright, the remaining combination of technologies, business models, and contracts appear inadequate to protect the laborious sweat in database production.

In summary, we conclude that there are two distinct processes and two distinct products wrapped within conventional use of the term "database." The process and product of selecting and arranging is protected primarily by copyright and by some combination of technologies, business models, and contracts. More ambiguous is the protection granted to

the process and product of gathering data. In the past, the creativity in databases was effectively protected by copyrighting the printed material. But modern technologies make it possible to separate the creativity from the facts. There is a question of whether copyright is the appropriate mechanism for protecting products of the selection and arrangement process, but that is the subject of a different thesis.

In the remainder of this chapter, we focus on the limited protection for data gathering. Assuming again the inevitably of government intervention due to both domestic and international pressure, the question is now, what measures can the legislature take to balance innovation and production in the gathering of data? Subsequent references to database protection in this chapter will refer exclusively to products of the process of data gathering unless explicitly noted otherwise.

4.3 How do we protect the data in databases?

Having identified what we are attempting to accomplish: balancing database innovation with incentives to produce, we turn now to consider possible mechanisms. The Policy Analysis reviewed a number of available public and private sector options. As noted earlier, in this chapter we focus on legislative options and adopt a more theoretical approach. In this section we introduce two different economic frameworks. These two economic models not only guide policy formulation but also sugest measures for evaluating success.

There are two different economic models which we might use to select and/or evaluate whether a specific legislative approach will fulfill the Constitutional mandate to "promote the progress of science and the useful arts (U.S. Constitution Art. 1 Sec. 8)." The general principle in both cases is to cast barriers to innovation as market failure. The first approach models the market for database production and innovation in the traditional Prisoner's Dilemma (Gordon 1992a). If both players shirk by focusing on creative selection and presentation rather than gathering data, there is no product. Successful interventions balance the payoffs to induce cooperative behavior. The second approach, entitlement theory, models the failure to innovate as the result of high transactions costs between parties who gather and parties who select and present. Legislative options take the form of "property rules" and "liability rules" that reassign the initial allocation of rights in an attempt to reduce transactions costs (Calabresi and Melamed 1972; Hardy 1996; Merges 1996; Perritt 1996)

4.3.1 Prisoner's Dilemma

The Prisoner's Dilemma is the classic single-stage, two-player simultaneous (static) game (Gibbons 1992). Although numerous variations have been applied to better fit the model to various scenarios, the original model has proven quite robust. Following Gordon (1992a), we apply the two-player framework to the policy challenge of inducing innovation in data gathering, selection, and arrangement in the context of the attribution problem space. We begin with a brief description of the Prisoner's Dilemma, review the general application of the game to intellectual property, and conclude with policy guidelines suggested by the game.

4.3.1.1 Prisoner's dilemma

The traditional prisoner's dilemma (PD) is told as a story of two criminals. A prisoner and his partner are imprisoned by the local sheriff for a crime they committed. Unfortunately, the sheriff has no evidence and must extract a confession in order to prosecute. The prisoner and partner are held in separate cells and prevented from communicating. Each criminal faces two choices. He can attempt to *cooperate* with his partner-in-crime and refuse to confess. In this case, the sheriff, lacking any evidence, can only imprison the criminals until their arraignment at which point they are both released and can divide the spoils. However, if the prisoner *defects* by offering to testify against his partner while the partner continues to keep silent, then the defector is immediately set free while the holdout is penalized both for the crime and for obstruction of justice. The payoffs are reversed when the prisoner cooperates but his partner defects. The final scenario is one where both criminals defect and implicate one another. In this situation, both prisoners are sentenced for the crime although neither faces obstruction charges. The payoffs are often drawn as a two-by-two matrix, mixing the payoffs of both players. We adopt the representation in Table 8.1 as possibly more clear (Tzafestas 2000).

Prisoner	Partner	Outcome	Payoff	Scenario
Defect	Cooperate	Partner is convicted of crime and obstruction	5	Temptation
Cooperate	Cooperate	Free at arraignment, split the booty	3	Reward
Defect	Defect	Both are convicted, no obstruction charge	1	Punishment
Cooperate	Defect	Convicted alone of crime and obstruction	0	Sucker

Table 4.1 Payoffs for one prisoner with respect to the behavior of his partner

Each player has two strategies. They can either *cooperate* with one another or they can *defect* against one another. Because the prisoners are not allowed to communicate, they effectively make their decision to cooperate or defect simultaneously. There are no appeals, no second chances, and no double jeopardy. Therefore, the game is only played once and constitutes a single-stage. Pivotal to the outcome are the relationship between the different payoffs and the single-stage, simultaneous (no communication) nature of the game.

From the table, it is easy to see that if the Partner chooses to cooperate, the Prisoner receives a bigger payoff by defecting. If the Partner defects, the Prisoner still does better by defecting. In other words, regardless of the Partner's behavior, the Prisoner always does better by defecting. In economic terms, the strategy of cooperation is strictly dominated by defection. Any time a total order exists where the "Temptation" scenario is unambiguously better than the "Reward" for cooperation which is in turn more valuable than the "Punishment" of being imprisoned which is better than the "Sucker" payoffs, one strategy is strictly dominated by the other (Gibbons 1992).

The single-stage nature of the game ensures that memory and the potential (threat) of future interactions do not color the outcome. Were players to play one another repeatedly, both strategies and outcomes would look quite different (Gibbons 1992).

Finally, when each player, acting alone, accounts only for his/her own interest, defecting is the rational strategy. However, it is clear that if the two players can reliably communicate and cooperate, both are better off. More significantly from a policy-maker's point of view, players maximizing personal incentives may not result in a globally optimal outcome. There are many applications of the PD, including the "free rider" problem facing public goods like information (Milgrom and Roberts 1992; Shapiro and Varian 1999). Defection leads to underproduction and lower overall social welfare. The Tragedy of the Commons, where farmers overgraze a public resource, is the classic application of the PD where overall social welfare suffers when players attempt to optimize personal profits (Gibbons 1992).

4.3.1.2 Prisoner's dilemma and intellectual property

The keys to the PD are the relationship between the payoffs, the single-stage nature of the game, and the lack of communication between parties. We consider each of these factors in the context of policies for intellectual property.

Applied to intellectual property, the two strategies of cooperate and defect are cast as producing or copying (Gordon 1992a at 863).[48] The payoffs for each strategy are most often hypothesized under the assumption that intellectual property is a public good (Gordon 1992a; Perritt 1996). The hallmark of public goods is their non-rival and non-excludable characteristics (Milgrom and Roberts 1992; Shapiro and Varian 1999). A non-rival good is one where use does not consume the resource. Unlike eating a meal, a person can read a book or listen to a song without exhausting the good for later reuse. Non-excludability is the property where multiple users can simultaneously enjoy the same good. Only one person can sit in an airplane seat on any given flight. Seats on flights are therefore excludable. However, every person on the flight can watch the same movie simultaneously.

The standard assumption is that because intellectual property is non-rival and non-exclusive, whatever the costs of production, the costs of copying (free-riding) are significantly lower if not zero. As a consequence, production as a strategy is strictly dominated by copying.

Perritt helpfully clarifies the standard assumption by offering one attempt at itemizing the costs associated with each strategy (Perritt 1996 at 278). Production involves: creation (cc), packaging for distribution (cp) (e.g. constructing a patented device or formatting a copyrighted work), marketing including billing (cm), and the standard marginal cost of producing an additional unit (mc). Copying involves similar marketing costs (cm) and marginal costs of reproduction which, assuming wholesale piracy, is the same as that of the producer (mc) (Perritt 1996 at note 63). Additional costs facing the copier are the cost of acquisition (ca) to find and access the intellectual property being copied, the cost of transformation to (re)package the good (ct), and the cost of legal liability (ll) in the event that the copier is sued. Because the producer's cost of creation and packaging are generally assumed to be much greater than the copier's costs of discovery, (re)packaging, and legal

[48]Perritt (1996), refers to copying as piracy.

liability, scholars (and producers) assume that production is dominated by copying. The condition is denoted: cc + cp >> cd + ct + ll (Perritt 1996).

The game is effectively single-stage because if both players elect to copy, there is no product to copy and no game. If one player elects to copy, the producer is driven from the market after a single stage and again, there is no subsequent game to play.

Merges et al. (1997)explain the single-stage condition by applying the PD to the economic model of a Bertrand, price competing duopoly. Assuming Bertrand competition, each player prices at marginal cost (Gibbons 1992). Where both players shirk, there is nothing to copy and each player experiences a loss equal to their investment as a copier. If both players cooperate, they split the market and each makes a modest profit. If one player cooperates while the second player shirks, competition again drives the price to marginal cost. However, the producer is then unable to recover her fixed costs of development, incurs her entire investment as a loss, and leaves the market (Merges et al. 1997).

The simultaneity of moves is similarly asserted by the public goods nature of intellectual property. Non-rivalness and non-excludability suggest that a potential pirate need not negotiate or communicate with a producer ex ante. The public never perceives scenarios where both parties choose to copy (defect) simply because the market never materializes. Given situations where the copier's costs are sufficiently low, both players shirk; the equilibrium outcome results in no production. From the perspective of our initial policy objective, to stimulate innovation both in data gathering and selection and arrangement, society is clearly worse off.

The implications for policy making are straightforward. Legal liability (ll), the remaining cost in Perritt's equation, represents the policy-maker's instrument for altering cost incentives. Where the differences in costs already approach zero, the need for intervention becomes small. To the degree that any intervention is justified, we move next to consider lessons from the PD for policy formulation.

4.3.1.3 Policy-making and the prisoner's dilemma

As observed by Gordon (1992a), the PD offers a number of lessons for the policymaker. It not only stipulates conditions under which intervention is justified, but also provides guidelines on appropriate action and metrics for evaluating success.

Conditions for intervention

The PD suggests four conditions for action: the presence of competition, the dominance of defection, the implicit desirability of cooperation, and the availability of viable interventions.

First, is there competition? If there is no competition then there is no game. There is no market failure. In pragmatic terms, the absence of competition means that defection does not result in a competing product that drives prices to marginal costs and precludes the cooperator recovering her costs.

Though seemingly straightforward, ambiguity in this condition arises from how broadly "the market" for a product is defined. Market definition is a significant issue for determining the presence of market failure inducing competition. Producers decide whether or not to produce (innovate) by identifying a set of needs (or uses) and a perceived set of customers by which to estimate demand (Pindyck and Rubinfeld 1992). For example, the market for lodging in Tokyo, Japan might be defined as all customers seeking a bed for the night. A competing product, by definition, addresses the same market, increases supply, and drives prices down. Hotels across the city compete in the same market.

If the customer pool required to recover costs is defined narrowly enough, there is room for other producers to enter the market and target a well-defined subset of customers. Differentiated products compete in only one segment of the original market and have a limited competitive effect on price (Pindyck and Rubinfeld 1992). Hostels, for example, focus only on those low-budget customers willing to share rooms and tolerate limited hours for entry and exit.

Producers who define their market broadly, however, are vulnerable to cream skimming (Tyson and Sherry 1997). Second comers (defectors in the PD) who target high-value customers can steal high profit margins intended to recover investments in innovation. The argument against competition in local telephony was that competitive access providers would steal high-value business customers in urban centers and leave the low-value rural residential populations underserved (Baumol and Sidak 1994). At the same time, differentiated products represent innovation and may produce products better tailored to users and uses.

Complementary products address the same set of customers but address a related need (Milgrom and Roberts 1992; Pindyck and Rubinfeld 1992). An increase in the price of a complement decreases the demand for the original good. Food services such as in-house restaurants (or dining halls in the case of hostels) complement the market for beds. Even complementary products are not without controversy. For example, is a Web browser a complementary product that increases the demand for operating systems, or is it essentially an integral component and thereby a competing product with the potential to ultimately drive down prices (U.S. v. Microsoft 2001)?

The second condition for intervention is the dominance of defection. Without government intervention, do the payoffs in the game suggest defection as the dominant strategy? More specifically, do the strategies and payoffs suggest a relationship where players, acting in their rational self-interest, find the temptation scenario most attractive followed by the reward scenario, punishment, and finally the sucker payoffs?

Implicit in labeling the dominant strategy as undesirable is the third condition: that cooperation (the reward scenario) is actually superior to the punishment scenario or the sucker payoffs. The standard PD explains behavior from the perspective of rational self-interest.

Taking into account only personal incentives, the reward scenario is clearly advantageous for both players.

In a metaphorical sense, however, it is not clear that inducing cooperation and allowing both criminals to walk free is desirable. Labeling the game the "prisoner's" dilemma highlights the policy-maker's need to consider overall social welfare. Is cooperation desirable where doing so puts two criminals back on the street? The Tragedy of the Commons (Gibbons 1992; Milgrom and Roberts 1992), a classic application of the PD, internalizes the policy-maker's challenge to consider overall social welfare in maximizing individual benefits. Other examples of incorporating overall social welfare include instances where one law preempts another as in the case of free speech pre-empting copyright (Gordon 1992a; Pollack 1999).

A final condition for intervention is the availability of viable mechanisms by which to intervene. Generally, are there mechanisms for altering the payoffs of different strategies? More specifically, in the context of specific production functions such as Perritt's cost equations for intellectual property, are there direct or indirect means for affecting specific costs?

Guidelines for appropriate action

Availability of policy as a condition for intervention points to the second lesson for policymakers. The PD offers guidance on appropriate intervention. In general, the formulation of the PD suggests that the policy-maker can either increase the costs of defection or decrease the costs associated with cooperation. To that end, production functions, as in the case of Perritt's equations for intellectual property producers and pirates, identify direct and indirect opportunities.

Direct intervention takes the form of increasing the costs of defection through legal liability. "To cure this situation, the law creates anti-copying rules in the form of doctrines such as copyright, patent, and misappropriation. These legal regimes alter the relevant payoffs (Gordon 1992a at 865)."

Policy-makers can also indirectly affect incentives by encouraging innovation to bring competing costs into greater alignments. From a cost perspective, the crucial indicator of market failure is not a high cost of cooperation or a low cost of defection. Rather, it is the difference between the two costs. As the difference diminishes, so to does the incentive to defect. Innovation can both decrease a cooperator's production costs and increase a defector's. For intellectual property, Perritt identifies a number of technological and market mechanisms like encryption that increase a defector's copying costs (Perritt 1996).

Evaluating success

A final lesson from the PD for policymakers addresses metrics for evaluation. Evaluation is notoriously difficult. Perhaps the only significant arbiter is any individual's subjective assessment of the health of the market. However, the PD, at least metaphorically, does attempt to offer a subjective metric. To the degree that one can identify distinct strategies, we

can ask whether the empirical outcome results in the reward scenario where players cooperate. More concretely, policy-makers are forced to identify explicit costs that they attempt to alter, whether directly or indirectly.

As a caveat, there are limits to employing any model, including the PD, as a normative policy guide.[49] The model draws its conclusions based upon a certain set of initial assumptions that may not inhere to particular markets. First, as alluded to above in discussing the desirability of mutual cooperation, the traditional PD does not aim to maximize overall social welfare. Second, it is not clear that the game is strictly single stage. More specifically, true competition rarely corresponds to an idealized Bertrand duopoly. Products may not be perfect substitutes and the game may persist over multiple periods. Third, the game is not necessarily static. Is there no communication between players such that their moves are virtually simultaneous? Intellectual property copiers might transact (e.g. license or otherwise contract) with intellectual property creators. Finally, not all participants may be fully aware of the costs faced by other strategies (incomplete information), and even with perfect information, players may not always act rationally. Non-economic factors may intervene (Gordon 1992a).

4.3.2 Entitlement theory

One possible limitation of the PD, that players communicate and possibly transact, is addressed directly in the second economic model we consider as a policy formulation guide. Entitlement theory stems from the seminal work by Calabresi and Melamed (1972) on ownership and rights related to physical property such as resource pollution, theft, or accidents. In this section, we begin with a description of the framework and then follow Merges (1994; 1996) and Hardy (1996) in applying entitlement theory to intellectual property. Unlike the PD, where policy lessons are drawn independent of the game, entitlement theory was explicitly formulated as a policy guide. Consequently, we discuss implications for policy-making when describing the theory rather than in a separate subsection at the end.

4.3.2.1 Entitlement theory

Entitlement theory is based on transactions cost economics, one view of how players maximize their personal utility. Based upon the theory as formulated by Ronald Coase, welfare maximizing behavior is defined in terms of the optimal allocation of resources (Coase 1988). Resource suppliers in a perfect economy costlessly locate and transact with consumers; these economic exchanges result in a socially optimal allocation of wealth (Merges 1994 at 2657; Milgrom and Roberts 1992 at 303). Furthermore, according to the theory, initial assignment of property rights is irrelevant because in perfect, frictionless markets, people with rights to resources willingly bargain with those who desire the goods. Unfortunately, markets are not frictionless. Transactions costs intervene (Milgrom and Roberts 1992 at 28). The transactions costs that preclude bargaining play the same market failure inducing role in transactions cost theory as the dominated payoff structure in the PD.

---

[49] Gordon (1992a) discusses the PD as neither necessary nor sufficient condition for action. She argues that the PD is insufficient in cases where the model assumptions break down. The PD is unnecessary in the sense that there may be non-economic justifications fro action or other incentives unaccounted for.

Entitlement theory suggests that government intervention reduces transactions costs through a combination of initial rights allocation and transaction inducing policy protections (Calabresi and Melamed 1972 at 1110). We examine transactions costs, the available policy interventions, and then the guidelines for intervention originally proposed by Calabresi and Melamed.

4.3.2.2 Transactions costs

The PD is a model for predicting behavior in a two-player, single-stage, simultaneous game. The metaphor of two prisoners is used to help illustrate the effects of strictly dominated strategies. To present entitlement theory, Calabresi and Melamed use the metaphor of transacting for use permits on a community river. The competing strategies in this case are to fish or to pollute (Calabresi and Melamed 1972).

Transactions costs are loosely divided in the economics literature into coordination costs and motivation costs (Milgrom and Roberts 1992). For hoteliers seeking guests and travelers seeking accommodations, the travel industry serves as an institution bringing sellers and buyers together. The costs of creating and maintaining the travel industry are coordination costs of the market for hotel rooms. Some coordination tasks are more costly than others. For a factory negotiating for the right to pollute a river, locating all of the affected fishermen competing for use permits may be as simple as posting signs for a public hearing or as costly as meeting every local resident to negotiate individually (Calabresi and Melamed 1972). In some cases, the task is so onerous (e.g. the cost of identifying all affected parties so high), that a market fails to form (Merges 1996).

Once buyers and sellers are paired, a successful transaction requires negotiating a price and executing (enforcing) the conditions of the bargain. Hotels post prices and travelers "bargain" by picking lodgings within their constraint set (e.g. cost, proximity, etc.). Reservations and deposits secure the transaction. Advertising, price discovery, and reservations systems are all motivation costs (Milgrom and Roberts 1992). Eliciting the value of polluting or the collective value of fishing untainted waters can prove more difficult than pricing hotel rooms. Given a lack of alternatives, strategic bargaining, where parties have an incentive to inflate or deflate the cost or value of polluting versus fishing can overwhelm interests in transacting. Enforcement (monitoring) costs can also be prohibitive. Detecting and verifying one factory's pollution is difficult if there is more than one factory or if an unrelated disease wipes out the fish population. High motivation costs can also preclude transactions (Calabresi and Melamed 1972).

4.3.2.3 Entitlements

In transactions cost theory, markets fail when coordination costs or motivation costs overwhelm the incentive to trade. The initial allocation and subsequent protection of entitlements are presented by Calabresi and Melamed as a means for tempering transactions costs (Calabresi and Melamed 1972).

Initial rights allocations affect the ability to achieve an optimal outcome in two ways. First, initial allocations are an incentive to trade because they establish the initial bargaining positions and effectively establish rights distributions in the event of failure to transact. Factories (or fishermen) know that if an agreement is not found, then fishermen (or factories) can simply enjoin (or take) the right (Calabresi and Melamed 1972). A second effect of initial allocations on optimal outcomes is in the presence of multiple equilibria. "What is a Pareto optimal, or economically efficient, solution varies with the starting distribution of wealth. Pareto optimality is optimal *given* a distribution of wealth, but different distributions of wealth imply their own Pareto optimal allocation of resources (Calabresi and Melamed 1972 at 1096)." Initial allocations are therefore a policy means for engineering the outcome.

Once established, Calabresi and Melamed identify three available mechanisms for managing transactions costs: Property rules, liability rules, and inalienability. Property rules[50] are strong entitlements and give preference to the owner (seller) both in negotiating and in the presence of a failure to transact. "No one can take the entitlement to private property from the holder unless the holder sells it willingly and at the price at which he subjectively values the property (Calabresi and Melamed 1972 at 1105)." When the rights holder sets the price, this is referred to as "individual valuation" (Merges 1996).

Liability rules, by contrast, give preference to the buyer. A liability rule is defined by "the right to take property with compensation (Calabresi and Melamed 1972 at 1105)." If the parties to a transaction fail to negotiate a price, buyers may simply take the good for a legislatively or judicially determined price. "[A]n external, objective standard of value is used to facilitate the transfer of the entitlement from the holder to the nuisance (Calabresi and Melamed 1972 at 1105)." Court determined reparations in the case of negligence (Calabresi and Melamed 1972) or compulsory licensing of intellectual property (Hardy 1996; Merges 1994; 1996) are two such examples. Price setting performed by other than the parties to the exchange is coined "collective valuation" (Merges 1996).

Inalienable rules are a third policy mechanism. Rather than promoting exchange, however, inalienability is an anti-trade mechanism. From an economic perspective, inalienability constitutes a legislatively or judicially determined finding that the costs of trade are socially unacceptable. As a consequence, "in some instances we will not allow the sale of the property at all, that is, we will occasionally make the entitlement inalienable (Calabresi and Melamed 1972 at 1106)." The sale of body parts is one example. However, as our initial presumption in pursuing the attribution-related problem area was data reuse and redistribution, we focus the remainder of our analysis on property and liability rules. Indeed when we apply the theory to intellectual property rights, we will see that other entitlements literature makes similar assumptions (Hardy 1996; Merges 1996).

---

[50] The use of the term "property rules" refer to entitlements and should not be confused with *intellectual* property rules (IPR) which are used in a distinct although related context. We discuss the relationship later. Briefly, some intellectual property rules, like patents and copyrights, are property rules in the entitlements sense. However, compulsory licensing is not a property rule.

To illustrate the interaction of initial allocation and protection mechanisms, Calabresi and Melamed turned to the negotiation between factories and fishermen over water resource rights (Calabresi and Melamed 1972). Where fishermen have the entitlement, which is protected by a property rule, the factory must pay whatever price the fishermen ask for the right to pollute. Should the factory hold the property rule-protected entitlement, fishermen must pay whatever price the factory seeks in order to stop the pollution. Where fishermen hold a liability rule-protected entitlement to the water resource, a factory can, by paying all fishermen a government determined penalty, pollute regardless of the fishermen's desires. Likewise, if the factory holds a liability rule-protected right to pollute, fishermen can pay the factory an externally determined price, thereby compelling the factory to stop polluting.

4.3.2.4 The entitlement framework for policy-making

Calabresi and Melamed build their framework by considering the interactions between coordination costs and motivation costs and then evaluating what combinations of initial allocation and protection are most appropriate.

Assuming some initial incentive to trade, initial allocations are assigned with an eye to minimizing motivation costs. In particular, the participants who are best able to estimate the true value of the right should receive the initial allocation. If discriminating among participants is not possible, then "the costs should be put on the party or activity which can, with the lowest transaction costs, act in the market to correct an error in entitlements by inducing the party who can avoid social costs most cheaply to do so (Calabresi and Melamed 1972 at 1097)." Essentially, the rights belong with the party who is best able to incur the costs of market creation.

Unlike the PD, where decisions are made to maximize personal utility, entitlement theory explicitly seeks to optimize more than economic efficiency. Recall that the economically efficient solution "is optimal *given* a distribution of wealth, but different distributions of wealth imply their own Pareto optimal allocation of resources (Calabresi and Melamed 1972 at 1096)." Consequently, entitlement theory attempts to consciously account for general social welfare through the initial allocation.

> [A] society which prefers people to have silence, or own property, or have bodily integrity, but which does not hold the grounds for its preference to be sufficiently strong to justify overriding contrary preferences by individuals, will give such entitlements according to the collective preference, even though it will allow them to be sold thereafter (Calabresi and Melamed 1972 at 1101).

Although the overall goal of intellectual property law is often described in allocational efficiency terms (i.e., to increase economic output by overcoming market failures associated with the public goods quality of creative works), there is often an undercurrent of concern with the distribution of resources (Merges 1994 at 2661).

Once rights are assigned, policy makers must then identify a corresponding rule to encourage entitlement transactions. The underlying assumption in entitlements theory is the belief that, where possible, markets are the ideal mechanism for eliciting value and setting prices. External valuations employed in liability rules have a tendency to under-value (Merges 1996). Consequently, where the motivation costs of valuation are high, property rights that rely upon markets to negotiate a price are strongly preferred (Merges 1996). Concomitantly, because property rights rely upon individual, negotiated agreements, property rights tend to apply best where parties face low coordination costs.

In situations where there are many suppliers and many consumers, where identifying parties to negotiate prices is difficult, liability rules are generally more appropriate. Potential for strategic bargaining, in particular, will favor liability rules. Because they rely upon external agents to set a bound on prices, liability rules often tend to apply best in situations where the motivation costs are low and courts or legislatures can be relied upon to arrive at reasonable prices (Merges 1994).

In summary, the entitlements framework, in general, favors liability rules where high transactions costs prevail. Property rules are favored where transactions costs are low. The caveat is high valuation costs, where market-oriented individual valuations that stem from property rules are preferred over government determined collective valuations. Calabresi and Melamed craft the framework by identifying some of the transactions costs, laying out a set of entitlements, and then creating a matrix to identify which entitlements apply in different scenarios defined by the presence or absence of particular transactions costs.

4.3.2.5 Entitlement theory and intellectual property

The key to entitlements theory lies in identifying both the presence and magnitude of transactions costs. We therefore consider transactions costs in the context of intellectual property. Moreover, we accept as a given the implicit initial allocation of rights to the original creator, author, or compiler. For intellectual property, then, the policy maker's challenge is to identify the appropriate rules to best facilitate socially optimal transactions.

Transactions costs in intellectual property

While intellectual property may be bought or sold like any other property, the public goods nature of intellectual property tends to exacerbate certain transactions costs. Recall from our discussion of the PD that public goods are characterized as non-rival and non-excludable.[51] Perritt (1996) draws the connection between these public goods characteristics and motivation costs associated with detection and enforcement of binding contracts. Because the cost of copying is low, information goods incur high transactions costs for monitoring and policing reuse and redistribution. "It would be extremely difficult in most cases for an intellectual property right holder to identify all potential infringers, and downright impossible to separate those who posed a serious threat of infringement from those who did not (Merges 1996 at note 23)." "In the [intellectual property] context, there is no smoky soot or wandering cattle

[51] See Section 3.1.2

to serve as an unambiguous marker, although a direct copy of an apparent feature may appear on the market in some cases (Merges 1994 at 2658)."

Merges identifies the same intellectual property transactions costs as Perritt and adds to them the additional observation that valuing intellectual property is often difficult. In particular, Merges comments on how intellectual property inherently builds upon prior work. Because of "the abstract quality of the benefits conferred by prior works and the cumulative, interdependent nature of works covered by [intellectual property rights] ...[valuation] is at least as great a problem as detection (Merges 1994 at 2659)."

Property rules versus liability rules in intellectual property

In IPR, the initial assignment of rights is implicitly to the creator, author, or compiler. The question is therefore how best to facilitate transactions given this initial assignment. As noted earlier, we follow Hardy in omitting inalienability as a policy alternative where our explicit purpose is to encourage exchange (Hardy 1996).[52] However, in addition to the standard property rules versus liability rules dichotomy, we present Merges' extension to the entitlements framework as applied to intellectual property. Merges introduces "private liability rules" in contrast to the government mandated price-setting of traditional liability rules (Merges 1996).

In the context of intellectual property, we can think of property rules as "ex ante" rights. "A property rule allows the right-holder to set her own asking price through ex ante negotiations when someone begins to interfere with the holder's activities (Merges 1994 at 2665)" We can contrast ex ante rights with liability rules or "ex post" rights. "[L]iability rules are best described as 'take now, pay later.' They allow non-owners to use the entitlement without permission of the owner, so long as they adequately compensate the owner later (Merges 1996 at note 17)."

Patents and copyrights are examples of property rules in the intellectual property arena (Hardy 1996; Merges 1996). A property rule in the data integration context would enjoin reuse or redistribution without the explicit permission of the rights holder. A liability rule would allow reuse or redistribution without any agreement. Compensation could be exacted ex post either through a standard legislatively or judicially determined fee schedule. In the absence of such a schedule, the rights holder could sue in court and exact a penalty (and possibly, by precedent, set a schedule for future instances.) Compulsory licensing schemes are examples of liability rules for intellectual property (Hardy 1996; Merges 1996).

To the original entitlement polarity between property and liability, Merges introduces private liability rules (Merges 1996). The defining characteristics of private liability rules are property rules for protecting entitlements but with prices set by collective valuation, as is the

[52] Hardy (1996 at 230-1) acknowledges the interesting dimensions but potential lack of relevance of inalienability when applying entitlement theory to intellectual property. Perritt (1996) and Merges (1994; 1996) implicitly make the same assumption by discussing only the contrast between property rules and liability rules.

case for standard liability rules. Collective valuation in the case of private liability rules is performed by a coalition of entitlement holders rather than by a government institution. Merges points to ASCAP (American Society of Composers, Authors, and Publishers) and BMI (Broadcast Music Incorporated) as examples of privately initiated and maintained Collective Rights Organizations (CROs) (Merges 1996). ASCAP and BMI represent groups of songwriters and artists as sellers to radio, television, and other entertainment outlets. Blanket licenses are issued, payments collected, and royalties distributed according to standard price schedules and remuneration schemes fixed by the CRO; monitoring and enforcement of license conditions are responsibilities of the CRO (ASCAP 2001; BMI 2001).

The entitlement framework and intellectual property policy

While intellectual property may be bought or sold like any other property, three factors add to their uniqueness. First, the public goods nature of intellectual property drives transactions costs up (Perritt 1996). Second, the repeated play characteristic of some types of intellectual property transactions can induce private institutional reform to drive transactions costs down (Gordon 1992a; Merges 1994; 1996). Third, information technologies tend to exacerbate particular transactions costs while tempering others (Hardy 1996; Perritt 1996).

First, the entitlements framework suggests that liability rules are most appropriate in situations where high transactions costs prevail. Many forms of intellectual property, including live and recorded works of authorship and performance, are characterized by markets with many buyers and sellers that tend to increase coordination costs. High coordination costs are compounded by the public goods nature of intellectual property. Motivation costs for monitoring and enforcement necessarily rise to compensate for the non-rival and non-excludable characteristics (Perritt 1996; Shapiro and Varian 1999). Finally, certain forms of intellectual property are especially susceptible to strategic bargaining. Blocking patents, in particular, can preclude innovation by denying inventors the right to improve upon novel inventions (Ginsburg 1990; Merges 1996; Paepke 1987; Reichman and Samuelson 1997). Liability rules overcome these high cost disincentives to trade. To capture the benefits of exchange, liability rules allow people to copy and negotiate ex-post.

High valuation costs, combined with the previously unaccounted for repeat-play dimension of transactions, tend to favor property rules. As noted elsewhere, the ease of appropriability, particularly in inventions, can significantly complicate intellectual property price-setting (Merges 1996). At the same time, the second factor, the influence of repeat play can have an impact (Merges 1994; 1996). The reasoning states that, where a strong preference for individual valuations (property rules) are counter-balanced by high coordination and enforcement costs which are compounded by repeated plays (liability rules), private collective rights organizations will emerge to fill the void. As noted also in our discussion of the PD and intellectual property, the influence of repeated plays on economic incentives is frequently overlooked (Gordon 1992a; Merges 1994). Collective rights organizations thus constitute a middle ground between property and liability rules. The blanket license provisions simulate

liability rules while collective valuation by agents for participants in the transactions (the CRO) proxy for individual valuation.[53]

Finally, information technologies both magnify and temper intellectual property transactions costs. Motivation costs associated with monitoring and enforcement rise. Digital technologies simplify the task of creating, while increasing the quality of, pirated works (NRC 1997a; Perritt 1996; Tyson and Sherry 1997). At the same time, coordination costs are reduced by electronic search and market-making tools that bring buyers and sellers together (Hardy 1995; Merges 1996; Shapiro and Varian 1999).[54] Greater access to timely information decreases information asymmetries between negotiating parties (Milgrom and Roberts 1992; Shapiro and Varian 1999). Digital data communications virtually eliminate delays in delivery (Hardy 1995). Technologies that compound enforcement costs can enhance the ability to monitor as well. Digital encryption, access control, and search technologies hold significant promise for reducing monitoring and enforcement costs (Perritt 1996). As a consequence, transactions costs for intellectual property, depressed by information technologies, will tend to favor property rules.

Applied to intellectual property, then, the entitlements framework follows the same general rule. High transactions costs favor liability rules and low transactions costs favor property rules. Policy makers should carefully consider the effects of repeated plays and information technologies, however. Repeated plays may stimulate private institutional formation (CROs) obviating the need for government liability price-setting. Information technologies can both depress or inflate existing transactions costs.

4.3.3 Relating the prisoner's dilemma and entitlement theory

Note the relationship between the PD approach and the entitlements approach. In the PD, market failure is portrayed as a failure to innovate represented by mutual defection. The incentives to defect can also be interpreted as the result of high transaction costs. In the general case of the PD, high transaction costs are associated with the inability to communicate (the simultaneous nature of the game) and the inability to make binding contracts (e.g. the prisoners could agree to cooperate but in a single stage game, were only one player to defect, the defector would walk and the cooperator would face the large penalty) (Milgrom and Roberts 1992). Perritt discusses the public goods nature of information (excludability and rivalness) as sources of transactions costs in the market for intellectual property (Perritt 1996). The role of government in the PD scenario is to introduce legal liability as an additional defection cost to balance out the disincentives created by high transaction costs.

---

[53] See Merges (1996) for observations on why private collective valuations are favored over government collective valuations.

[54] Hardy (1995) discusses the effects of search technologies. Merges (1996) questions whether electronic marketplaces could replace the need for physical markets altogether, at least for information goods. The creation and subsequent implosion of a number of on-line markets (paper exchange, steel exchange, chemical exchange) suggest both the potential and the limitations of on-line markets at least for physical goods.

In summary, the PD models market failure that results from misaligned incentives between competing strategies that result in sub-optimal outcomes. Policy lessons are directed at realigning those incentives. Entitlement theory models the market failure that results from high transactions costs. The theory presents the initial allocation and subsequent policy protection of entitlements, as a means for overcoming failure inducing transactions costs.

4.4 Protecting data: databases as a unique form of intellectual property

In our presentation of the two different normative frameworks of PD and entitlements, we described each framework, the general application of that framework to intellectual property, and the attendant policy implications. In this section, we now revisit each framework in the specific context of the two processes and products associated with databases. Our conclusion is that the differences between databases and other, more familiar types of intellectual property (e.g. music, books, devices) suggest the need for a novel approach.

4.4.1 Prisoner's dilemma and databases

As with its general application to intellectual property, modeling the database market as a PD requires simulating the relationship between the payoffs, the single-stage nature of the game, and the lack of communication between parties. For databases in particular, we retain the single-stage, simultaneous interpretations of the game. However, the two distinct processes of the database market challenge conventional wisdom regarding the strict dominance of defection induced by the payoffs from public goods.

We model the two strategies in the database market as (1) gathering, and (2) selecting and arranging. As in the PD, is mutual defection where both players choose to select and arrange the strictly dominant strategy? Are the players (and society overall) better off under cooperation where both choose to gather? To answer these questions from the PD perspective, we need to consider both the costs and the payoffs associated with each strategy. We review costs using Perritt's cost model and payoffs by revisiting the market models from the Policy Analysis.

4.4.1.1 Costs in the database dilemma

To analyze the payoffs, we return to Perritt's characterization of the cost structure. Defection is induced when cc + cp >> cd + ct + ll. We assume for the moment that those calling for strong property rights in data gathering are correct and that ll is essentially zero. We focus instead on the remaining cost variables for both strategies.

Consider the defector's cost of transformation (ct). As suggested in the policy analysis and indicated in Section 8.2 describing the two processes of gathering and selection and presentation, new information technologies can dramatically lower (ct). A new presentation can be rendered with a single style sheet (Grosso and Walsh 2000; Raggett 2000; Walsh 1997). However, as also noted earlier, the true cost of transformation is not captured by the script, which specifies a style sheet. Rather, the true (ct) needs to reflect the creative work in

designing an interface for specific users or uses. This is not to suggest that (ct) is always high or that there is no threat from pirates who unimaginatively craft trivial changes. Rather, it is a caution to the policy-maker who might erroneously equate the simplicity of coding a stylesheet with the creative cost of transformation.

Compare the defector's cost of transformation (ct) to the cooperator's cost of packaging (cp). Note that the same tools that enable follow-on defectors to (near) costlessly craft execute presentation styles apply also to initial data gatherers. The true (cp) captures the creative considerations in identifying user populations and their respective needs. Does the defector have a cost advantage? Is (cp > ct)? Almost certainly. Many user populations may share overlapping interests enabling a follow-on data integrator to learn from those who came before. Our contention is first, that the difference in costs may be less than imagined and, more significantly, that it is this very learning that is the essence of what it means to "promote progress."

Defectors incur a cost of acquisition (ca) in lieu of creation costs (cc), according to Perritt's analysis. Evolving information technologies like the Web undeniably decrease (ca). Decreasing search costs is their intent if not their effect (Bailey 1998).[55] However, as Perritt also observes, both new market models and innovative technologies are evolving to help control the non-excludable and non-rival public goods characteristics of all information goods (Bloomberg 1996; Perritt 1996). Data gatherers in particular have long used market models successfully to control access (Perritt 1996). The market effects of the legal trials and tribulations of the on-line music industry testify to both the angst and innovation sparked by the specter of widespread reuse and redistribution (Hu 2000; MP3.com 2000; RIAA 2000).

However, information technologies also impact the data gathering cooperator's cost of creation (cc). The same tools that data integration defectors use to search for existing databases on the Web are available to data gatherers. Information technologies can significantly decrease the cost of database creation. "[O]ver time, the shift toward electronic databases may well reduce some of the upfront costs of entry, as the prices of hardware, software, and communications technologies continue to fall (Tyson and Sherry 1997 at note 32)." Perritt (1996) cites the example of creating a new, domain specific Web directory and the ease with which a pirate can copy the links to illustrate how new technologies decrease a pirate's cost of acquisition. However, he neglects to observe that the Web, which allows others to "steal" the new directory by framing or redirection, also supports search tools that greatly reduce the costs of creating domain specific directories in the first place. On-line filing, mark-up technologies, and text processing are decreasing the costs associated with legal electronic bankruptcy filings while creating an on-line database of cases (Markon 2001). Zagat Survey LLC is a leading international restaurant review guide. Expenses for producing regional restaurant guides include "printing and mailing surveys and then retyping user comments into a database.... 'When someone votes on the Internet, they are doing the data processing for us,' says Mr. Zagat, saving the company about $10 apiece for longer surveys

[55] Information overload is a classic refrain regarding today's Web (search papers)

(Shrager 2001)." More significantly, the attribution technologies from Part 1 as well as other innovations in data quality are aimed directly at the problem of data maintenance. For example, data quality improves while costs of data verification decrease because data integration technologies that remove human intermediaries can eliminate transcription errors (Huang, Lee, and Wang 1999).

4.4.1.2 Dominant strategies in the database dilemma

The PD perspective reveals that, at least for some market models, mutual cooperation is not necessarily optimal. Rather, cooperation in the colloquial sense may instead involve parties who gather data working in concert with those who select and arrange. Data gathering may be viewed as an "input" to the process of selection and arrangement much as Merges models ASCAP or, more generally, images, music, and video as inputs into multimedia products (Merges 1996).

> [D]atabase producers may negotiate with potential competitors who are interested in licensing a database and incorporating it in a competitive product. The database producer will try to negotiate a price that reflects his assessment of the value of the resulting competition product in the marketplace and the likely decrease in revenue from the original product (Tyson and Sherry 1997 at note 33).

Consider again the market models from Chapter 7. Recall that, depending upon the market models, certain approaches *benefit* from widespread (re)distribution of data. Thus, there may, in some market models, be an incentive to redistribution. The existence of mutual gains from trade suggest an incentive to transact. Whether and when such conditions exist is the subject of transactions costs and entitlement theory.

4.4.2 Entitlements and databases

To view the database market through the entitlement lens, we identify how characteristics of database products and processes impact coordination and motivation costs. The entitlement at issue is the right to creatively select and arrange an existing data set. Distinctions between intellectual property in general and the commercial database market in particular again challenge our initial intuition. For the specific purpose of transactions to support database innovations such as reuse and redistribution, property rules and even private liability rules may prove ineffective.

4.4.2.1 Coordination costs

First, we consider the coordination costs associated with a market of data gatherers and data arrangers. We begin with a remark on the impact of information technologies on a market for entitlements in selection and arrangement, examine the size of the market with respect to numbers of producers and consumers, and question whether costs are compounded by repeat transactions.

As a digital, network accessible product, electronic databases are a textbook example for which information technologies significantly reduce search costs. As noted in the PD analysis

above, information technologies reduce the defector's cost of acquisition associated with selection and arrangement.

The significance of the reduction, however, is directly dependent upon the size of the problem to begin with. While the commercial database market is quite large (Gale Research 1999; Tyson and Sherry 1997), that market is heavily differentiated (NRC 1997a; Tyson and Sherry 1997). As a consequence, from a coordination cost standpoint, pairing buyers and sellers, managing multiple customers, and managing multiple suppliers are all limited.

Whether there is significant competition within a niche is the subject of some controversy, which we address below. However, there is little disagreement to a characterization of largely domain specific producers and consumers with a manageable number of suppliers. Even in examples of competitive niche markets cited by proponents for strong protection, there are at most a handful of significant competitors (Tyson and Sherry 1997).

Moreover, the domain-specific nature of the overall market de-emphasizes the costs of coordination between producers. With libraries and universities as notable exceptions, demand both drives and reflects market differentiation. For example, even if there is competition in the production of financial data sources, customers of financial data will rarely be interested in purchasing the latest genomic database for commercial pharmaceutical research and vice versa. The contrast with collective rights organizations, that generate significant fees from blanket licenses, seems clear (Besen, Kirby, and Salop 1992).[56]

The market for commercial databases is also characterized by repetition. Merges notes that, in general, "[I]nput markets are notable especially for the repeated costs of locating right holders and negotiating individual licenses (Merges 1996 at note 62)." Databases in particular, because of maintenance and updating, engender high transaction repetition. For example, Zagat updates the data in their restaurant guides several times per year (Shrager 2001). A travel information integrator who makes use of regional restaurant reviews would want to consider following suit.

Despite repeated transactions, which inflate coordination costs, we argue that the effects of information technologies, combined with few producers, and highly differentiated markets, ultimately reduce costs. However, the combination of limited supply and narrow markets may indicate a vulnerability to strategic bargaining, which we turn to next.

4.4.2.2 Motivation costs and strategic bargainning

Though the costs of matching buyers and sellers may be low, database markets may prove different from other IP markets, like musical recordings, in their vulnerability to strategic

[56] To be sure, CROs do more than coordinate the transactions costs associated with blanket licensing (Merges 1996). Such activities are undeniably a significant part of their function, however, and the absence of such demand may decrease the incentive for independent evolution of a CRO in the commercial database market.

bargaining. The heart of the problem lies in the number of competitors and the appropriability of selections and arrangements.

Though there seems little disagreement on the differentiated nature of the commercial database industry, the degree of competition within each niche is hotly contested (NRC 1997a; Pollack 1999; Reichman and Samuelson 1997; Tyson and Sherry 1997). From a strategic bargaining perspective, however, the nature of the (debated) competition is not articulated. Applying attribution composition defined formally in Part 1 of this thesis, we conclude that much of the competition is based not on the data but on the selection and arrangement of that data. Consider the financial data industry, suggested as an exemplary, competitive, commercial data market (Tyson and Sherry 1997) . Financial information services certainly include proprietary analyst reports and market summaries. However, much of the *data*: stock prices, sales figures, earnings reports, derive from the *same* sources.[57] Competition exists because different providers, understanding the specific needs of energy traders versus analysts in currency markets select and arrange data to accommodate and optimize tailored needs.

Examples and anecdotes of competition cited by proponents of strong database protection reinforce the vulnerability to strategic bargaining. Proponents observe that, "with the profusion of freely available information (for example, on the Internet) and powerful computers and computing tools, database makers face competition worldwide from competitors and end-users alike." Yet the irony here is that such competition, to the degree that it exists, depends upon the ease of transacting the entitlement to select and arrange. "The data is the data. We believe the difference is the accuracy, timeliness, ease of use and search, and other feature capabilities we can provide (Tyson and Sherry 1997 at note 36)."

Note that the hazard in negotiating *with a competitior* over the right to compete is exemplified in *Feist*. Feist attempted to negotiate for the license and indeed acquired licenses from all other carriers in the regions for which he was creating an integrated directory. *Rural* refused to license, at least in part, explicitly because of its interest in entering the market itself (Feist v. Rural 1991).[58]

In addition to competition, the difficulty of valuing the entitlement increases the vulnerability to strategic bargaining in two ways. First, in speaking of patents, Merges notes the inherent difficulty of eliciting value in a follow-on product. How much of the value is due to the patented input and how much of the value is novel (Merges 1994)? In the database context,

---

[57] Stock prices come from the particular markets in which each stock trades. Earnings reports sales figures, and related data are collected as a Securities and Exchange Commission (SEC) regulatory requirement. Sole source monopoly providers like the stock exchange and data collected by government mandate are outside the explicit scope of this thesis. It should be noted, however, that it is a specific fear of some financial data services providers that strong data protection would confer a strategic bargaining advantage to sole source data suppliers like the financial markets (Bloomberg 1996).

[58] Tyson notes that instances where a monopoly provider refuses to license might best be dealt with independently under essential facilities doctrine (Tyson and Sherry 1997).

how much value is in the underlying data and how much value is in the selection and arrangement?

Second is the vulnerability of selections and arrangements to appropriation. A selection or arrangement, particularly one developed by focusing on the unique needs of a particular market segment, is easily duplicated once revealed. Merges draws an analogy to Arrow's paradox: "if in trying to strike a deal [a person bargaining for the entitlement to select or arrange] discloses her idea (e.g., the technology she invented), she has nothing left to sell, but if she does not disclose anything the buyer has no idea what is for sale (Merges 1994 at 2657)." While we know from *Feist* that the originality in selections and arrangements are copyrightable, a creator would have a difficult time defending the idea absent manifestation in an application which requires the entitlement.[59] Without creating the application first, the originality might never see the light of day.

4.4.3 Protecting data

In summary, our position is not to suggest that there are no differences in cost or that the market is not competitive or does not require protection. But competition in the database industry lies not in database production; competition lies in creative selections and arrangements. This is the innovation that we wish to foster.

To be sure, if no one gathers, there is nothing from which to select or arrange. We do not claim that there is no difference in costs between gathering versus selecting and arranging. We do not suggest that data gatherers face zero free-riding potential. Our contention is that the difference between the costs is arguably far less extreme than is commonly asserted.

Meanwhile, as technology makes new creative selections and arrangements possible, the threat to innovation shifts from the market failure of mutual defection to the market failure from strategic bargaining – failure that precludes new selections and arrangements from seeing the light of day. Strong property rules reinforce the tendency to failure. While strong property rules, combined with repeated transactions could induce leading to private liability rules, CROs may not be strong enough to overcome the strategic bargaining issue that a true liability rule is intended to resolve (Merges 1996). Moreover, for CROs to form, you need a critical mass of transacting parties (Ginsburg 1990). As argued above, there is reason to believe that the differentiation that characterizes database markets may not satisfy this threshold.

The challenge is therefore to overcome the tendency to strategic bargaining while balancing the failure inducing difference in costs between data gathering and selecting and arranging.

---

[59] There is an obvious opening to further analysis along the lines of protection of selections and arrangements under the performance copyright rather than strictly as a musical or written compilation. See discussion under future work.

4.5 A Federal statute of misappropriations in databases

We propose a Federal statute of misappropriations for databases as the most appropriate legislative intervention to balance the competing interests of the different stakeholders while pursuing the legislative mandate to promote progress. We discussed misappropriation doctrine as a general framework in the policy analysis of Chapter 7. We do not attempt to generalize and determine whether the doctrine is applicable to other intellectual property domains. Here, we focus on dimensions of a misappropriation doctrine specifically aimed at the innovation represented by the market for commercial (re)use and (re)distribution of data in commercial databases.

In this section, we follow Paepke (1987) in defining misappropriation operationally as a set of tests, which could serve as either a policy or judicial guideline for invoking a legitimate claim of misappropriation. Note that the conditions work in concert. A successful claim should satisfy all of the conditions. Possible remedies are suggested at the end.

4.5.1 Significant investment on the part of the creator

The producer needs to invest in order to claim protection (Paepke 1987 at 70). Recall also Perritt's cost model (Perritt 1996). The problem is not solely that the copier has a low cost of production. The issue is whether the difference between the producer's costs and the integrator or innovators costs would permit a second comer to sustainably price below the original producer.

The question of significant investment is particularly pertinent because of markets where the data gathered is ancillary to the good or service. Ignoring the federal monopoly dimension of "Rural Telephone," Rural would have gathered directory information as a function of its billing records. The gathering of data for the database would not justify a "significant investment (Feist v. Rural 1991)." Recall that the purpose of the claim is to promote and/or protect the incentive to invest in creation. Rural Telephone would have created a database of names and numbers regardless of whether Feist had attempted to compete. (An open question is whether Rural would have bound and published the telephone book and so we consider not only the difference in costs between producer and pirate but also additional factors identified below). By contrast, ProCD was not a telephone company and did not create telephone books as an ancillary good or service (ProCD v. Zeidenberg 1996). At least on its face, ProCD would have a greater claim to "significant investment" than Rural.

Note also that significant investment (and subsequent grounds for a misappropriation claim) can apply to both the process of data gathering and the process of data selection and arrangement. Patent law has a similar provision related to non-obviousness (Merges et al. 1997)

4.5.2 Appropriation by the defendant

A second condition for a misappropriation is actual appropriation by the defendant to a claim. Substantial similarity is not grounds for action. However, the question is whether the

integrator free rides on the plaintiff's investment in data collection and maintenance. It is important to note that the claim is based upon the plaintiff's investment and the defendant's use of the database in lieu of their own investment.

There are two significant dimensions to our appropriation condition. First, consider that the condition focuses on appropriation and not *when* that appropriation occurs. The implication is that whether one pre-fetches and warehouses or whether one queries in real-time, to the degree that data is appropriated, a claim is possible.

The second dimension of the appropriation condition concerns *why* data is appropriated. In particular, *why* does not matter. Integrators may act on behalf of a specific customer (i.e. as the agent for a client) or in anticipation of more efficiently serving future clients. If she makes use of someone else's data, she raises the potential for action.

Contributory liability is a subtle distinction in the appropriation. What of the integrator that creates a tool to aggregate data from a number of prespecified sources? For example, suppose that Zeidenberg has access to any number of directories, each in its own particular physical, logical, and conceptual arrangement; Zeidenberg chooses to develop a tool tailored to reusing and redistributing data explicitly from some prespecified subset (e.g. ProCD). Whether Zeidenberg develops the tools and sells the tools to individual users (personal use) or creates a service to mediate requests from users, we borrow from the copyright literature to conclude that in this circumstance, Zeidenberg bears contributory liability (Sony v. Universal 1984).

By contrast, consider the inventor who develops general theories and tools for aggregation. As companies become increasingly global, knowledge management within institutions is a burgeoning field. Much of the knowledge within any particular enterprise is captured within internal documents stored in heterogeneous fashion. Integration for knowledge management is only one of many possible markets for integration technologies and services (Lee et al. 1999). What then of the user who configures the tool to misappropriate? Again borrowing from *Universal*, if there are substantial non-infringing uses, the integrator does not bear contributory liability.

The distinction may seem arbitrary but is quite significant. If a defendant to a misappropriation claim creates a tool that has no other purpose than to appropriate an explicit target's data, then whether the defendant creates a service that responds to user queries or sells the tool and individual customers infringe, the defendant bears some measure of liability.

The defense against this condition is independent creation. Substantial similarity is not by itself sufficient for establishing appropriation. Again borrowing from copyright, independent creation is permitted. In database terms, if a competitor independently gathers data from base sources or collects specific user requirements and preferences to select and arrange, the resulting product does not constitute (mis)appropriation.

4.5.3 Use in competition with the plaintiff

If there is no competition, then there is no diminution of incentive and there is no grounds for a claim (Gordon 1992a). The proposition is that an inventor invents and produces with a particular business model in mind. "The inventor depends on a return on his investment from the product he develops, not from unanticipated off-shoots into other markets (Paepke 1987 at 72)." Therefore, in the context of database production, the deciding factor is whether, for the market defined a priori, whether the initial producer can recover her investment in data gathering.

One difficult question that arises is whether, in speaking of "return on investment," one includes "potential markets." What then of the producer who claims that they were "intending" to pursue a market and simply had not yet done so? The original producer may have been waiting for revenue from an initial market to generate sufficient capital to pursue a secondary market. Perhaps because the firm was just starting up they lacked sufficient human capital or other resources to pursue multiple markets in parallel and so had embarked upon a plan of sequential build-out. Alternatively, the intended market may not prove sufficient to justify continued investment, but the expansion to some unanticipated market may, in combination, provide sufficient return. Such a claim requires careful balancing against the initial condition of significant investment. "[T]he element of use in competition with the plaintiff is intended to focus the misappropriation remedy on free-riding that discourages efficient investment in research and development (Paepke 1987 at 72)."

Rural, for example, claimed that they were intending to (eventually) pursue the market targeted by Feist. However, even excluding government mandate, it is not clear that Feist's market was necessary to induce Rural to produce the original database. Even assuming that it had made a significant investment, the case does not indicate that Rural demonstrated any effort to develop the market before Feist's entry (Feist v. Rural 1991). In contrast, if a company could demonstrate, perhaps through documentation, prototyping, and other development signals, that they were intending to pursue a market that had been taken by a follow-on copier, the initial producer would have grounds for a claim.

There are a few additional factors that are often considered in proposals for misappropriation statutes. We consider them here but also indicate why we believe these additional considerations may be subsumed by the factors noted above.

4.5.4 (Lack of) signficant investment by the copyist

It has been pointed out that competition is driven, in part, by a level playing field that presents all parties with the same, initial transactions costs (Perritt 1996). Therefore, an additional condition sometimes raised to defend against a misappropriation claim is the demonstration of significant investment by the second comer (Gordon 1992a; Perritt 1996; Reichman and Samuelson 1997). If the copier incurs significant investments and therefore faces equally high production costs, then the second comer, competing in the same market, would have similar cost recovery constraints on pricing and competitions.

However, such an argument seems redundant. Whether he invests a great deal or whether he invests nothing, if a copyist does not compete with the provider, then from the standpoint of innovation, he in no way reduces the producer's incentive. At the same time, a new application (effectively, an innovation) is developed. There appears limited reason to object. "The inventor depends on a return on his investment from the product he develops, not from unanticipated off-shoots into other markets (Paepke 1987 at 72)."

Conversely, consider the copier who competes with the original producer. If the copier invests nothing (i.e. essentially competing in the same market with an identical product), then he can undercut the producer, justifying the misappropriation claim on competition alone. If the integrator or copier invests a great deal, thereby driving up her costs such that both the original producer and the second comer charge equivalent prices, the investment does not excuse the free-riding. Assuming even minimally intelligent investment, the copier, having avoided initial database creation costs, should have developed a superior product. This is the very investment and competition that we seek to encourage. Yet, if the new product completely displaces the producer's initial effort without any compensation, the initial incentive to produce is lost. The misappropriation claim would be based upon the appropriation, not the amount of investment.

The logic behind misappropriation as a liability rule is to encourage innovation through data reuse and redistribution. Suppose one develops a new application by tailoring an interface or providing enhanced data manipulation tools for a narrow market. The inventor seeks to license and the initial creator refuses. A liability rule allows the second inventor to reuse today at the cost of a penalty tomorrow. Is society better off? What is the value of a new interface or new tools? The issue is not how much investment was required to develop the new interface (i.e. it could be non-obvious but still cheap to produce). As noted by Merges (1996), the product of data gathering is now an input to a new product.

4.5.5 Appropriation of a significant amount

There is no small disagreement over how much of an appropriation is required to justify action (databasedata.org 1999). There appears a large continuum between one or two rows, (largely uncontroversial under fair use) and wholesale copying of an entire database (again, largely uncontested as a clear cause of action).

It seems that the critical question, however, reduces again to the issue of a producer's incentives. The quantity of data extracted says little about the degree to which it will impact the producer's incentives to produce. As an extreme example, consider a comprehensive, national telephone directory. If one were to copy the entire directory and use the contents as filler to manufacture doorstops, the use would likely not prejudice the initial creator's incentive to produce.

It should equally be noted that even a relatively small extraction, as measured by quantity, can directly impact the original producer's market. For example, a database producer might

compile a comprehensive listing of all commercial airline flights and schedules in the United States. As measured quantitatively, the subset of all flights along the Northeast Corridor between Washington, D.C., New York City, and Boston is a relatively small percentage of the total database. Yet this smaller subset could prove a significant competitor because a significant percentage of total air traffic is concentrated along this corridor. More generally, we might consider the issue of product bundles and their effect on market differentiation (Bailey 1998).

Determining what constitutes a "significant" amount independent of market competition is also problematic. Consider the case of an aggregator who gathers data on the behalf of specific clients. Each individual user may extract only an "insignificant" amount, but the net effect is to diminish the overall product. At the extreme, consider the case of an integrator that queries the initial producer in "real time" so never actually warehouses and extracts the data. Yet the cumulative effect is of a significant appropriation.

Moreover, extracting a "small" amount does not guarantee that *use* is limited to a small amount. Recall the discussion of negation from Part 1 of this thesis. To determine that a value is *not* in a particular table requires evaluating every row in the table.

What is or is not significant is problematic to determine. We therefore propose a different guiding principle. Whether the extraction is sufficient or not, the disincentive to the producer would stem from potential lost revenue. Even a small amount of data could be worth a great deal. The principle cause of action again, it seems, is the competition and not the amount of the extraction.

Having identified the conditions for a valid claim of misappropriation, we now turn to the question of remedies associated with a successful claim. The relief associated with liability rules is typically some combination of monetary penalty (e.g. royalties) and injunction (e.g. prohibiting outright or delaying the appropriation) (Paepke 1987; Reichman and Samuelson 1997). To balance these measures in a database misappropriation statute, we return to the two theoretical frameworks outlined earlier.

4.6 Misappropriation relief: the policy proposal in theory

Misappropriation supports two avenues for relief. There are fees or penalties associated with an ex post assessment of market impact and injunctions against everything from current or future appropriations to sales of products resulting from such misappropriation. In this subsection, we consider the problem and these methods of relief in the context of the two frameworks.

4.6.1 Misappropriation and the Database Dilemma

Intellectual property, like the traditional Prisoner's Dilemma, presents the problem of incentives that encourage mutual defection rather than more desirable cooperation.

Differences in costs of piracy versus production threaten the economic viability of intellectual property producers in general.

For commercial markets in data, however, the cost imbalance might not prove so large as otherwise assumed. As costs align, the incentive to defect decreases, suggesting that defection is less likely. Decreases in the cost of data collection may level the costs of piracy and reformatting with those of initial gathering that is targeted and tailored to a niche application.

Second, the highly repetitive nature of transactions in markets where data is an input also deviates from the standard PD model. As noted earlier, in repeated games, rational behavior induces cooperation that might obviate the need for further, statutory intervention (Gibbons 1992). In the database context, where data serves as an input to follow-on integration and innovation, the need for repeated updating (NRC 1997a; Tyson and Sherry 1997), even for products in direct competition with the original gatherer, suggests that the second-comer must ensure that the original gatherer captures sufficient return to induced continued production. Sufficient return might come from licensing or by segmenting the market to minimize direct competition between initial gatherer and follow-on producer.

In instances where costs remain skewed, inducing defection, misappropriation exacts an ex post liability cost from the pirate. Because valuation is difficult and follow-on producers have a disincentive during bargaining to reveal their innovations for fear of appropriation, ex post liability proceedings allow Courts to observe the marketplace as an indicator of both costs and value. Penalties in the form of profits might encourage ex ante bargaining from both parties to avoid the excess enforcement costs of litigation.

Injunctions, from the PD perspective, balance costs by forcing a loss of the pirates' initial investment outright. At first glance, enjoining sales of appropriated data may appear to offer little relief to a data gatherer. Hearkening to the limits of contracts, once the data is released, the value is arguably unrecoverable (Elkin-Koren 1997; Hawkins 1997; O'Rourke 1997; ProCD v. Zeidenberg 1996; Tyson and Sherry 1997). For database markets sensitive to timeliness, however, injunctions against future appropriation do place a bound on the vulnerability of first movers in data gathering.

4.6.2 Misappropriation and database entitlements

In the entitlements case, we saw that difficulty in valuing both initial data gathering and creative selection and presentation (i.e. what value stems from the data and what value stems from the arrangement) contributes to high transactions costs. Ease of appropriability compounds the problem of valuation in ex ante bargaining because second-comers have an added disincentive to reveal their ideas. That follow-on innovation may compete directly with the initial gatherer is an inducement to strategic bargaining. Some first movers may choose not to negotiate altogether. All of these factors contribute to market failure.

Liability rules allow ex post pricing to proxy for market negotiations that otherwise do not take place. In addition, as an inducement to bargaining, liability rules address the problem of appropriability for data selections and arrangements. Second comers who fear revealing their ideas in a priori bargaining know that in the absence of agreement, the innovation may still see the light of day. Finally, liability rules may indirectly alleviate some of the difficulties posed by database valuation, albeit indirectly. Additional information about the value of the product (if not information about the value of the data versus value of the selection and arrangement (Merges 1994)) is revealed by allowing the innovation to see the light of day and enabling a market to form.

In some instances, permitting second comers to appropriate will skew the balance in transactions costs too heavily towards data selection and arrangement. To correct the imbalance and induce second comers to bargain, misappropriation substitutes a high transaction cost of enforcement. Injunctions and penalties could not only penalize the second-comers initial investment but also exact the cost of ex post valuation for enforcement entirely from the follow-on producer. The threat of litigation to determine the ex post cost may therefore serve as an incentive to reach a transaction in advance (Merges 1996 citing Ayers and Talley at notes 21,22).

That a premium is placed on progress of science and the useful arts does not suggest that data gathering as an input has no value. If for no other reason, as noted in the PD, without compensation, there is no incentive to gather and no basis from which to "progress" from. At the very least, misappropriation can separate motivations for strategic bargaining. Injunctions simulate the refusal to bargain. Injunctions would be granted only in instances that prejudice initial incentives for creation.

4.7 Objections

There are a number of possible objections to the rule, and we seek to address some of them here.

4.7.1 Interference with private bargaining/incentive to bargain

A primary objection to any liability rule is the observation that liability rules remove the incentive to bargain for a price above the liability price (Merges 1996). Liability rules effectively preempt any attempt at allowing the market to determine a price because the data integrator has no incentive to bargain above a legislatively established liability price.

However, there are circumstances under which liability rules can induce bargaining by both parties (Merges 1996). Data gatherers face the threat of appropriation. The second comer may simply take the product. Knowing that the alternative is an externally determined price, liability rules can overcome impediments like strategic bargaining and refusal to trade by data gatherers.

From the integrator's perspective, we recognize that different collections of data inherently have different values. As an innovator, the integrators market may be untested and unproven. However, the liability rule can induce second comer's to bargain by enabling the integrator/innovator to reveal their innovation. There is a fear that the rights holder could steal novel selections and presentations by looking at an innovator's ideas, refusing to license, and then creating similar services. Appropriation ensures that in the event of a failure to license, the initial integrator may still bring their idea to the marketplace.

4.7.2 Bias the market in favor of second comers

Assuming that both parties agree to bargain, more significant is the danger that externally (judicially, legislatively, or administratively) determined liability prices skew the market. Second comers could refuse to reveal valuations above externally determined prices because they can always pay the liability price by simply misappropriating (Merges 1996).

Under some circumstances, price schedules can induce faithful valuations. Where a predetermined liability price is both above the value of some and below the value of others seeking to bargain with the rights holder, liability rules can lead parties to reveal their true value (Merges 1996 citing Ayers and Talley at notes 21,22).

However, recognizing the limits of scheduled prices, we aim to induce more faithful bargaining through case-specific ex-post penalties. In addition to the aforementioned danger of biasing negotiated prices, predetermined price schedules are subject to lobbying and are inherently inflexible over time (Ginsburg 1997; Merges 1996). Therefore, rather than codifying a price schedule, we rely upon courts to determine case-specific penalties. Ex-post penalties could prove costly to innovators in multiple ways. Injunctions could either bar the innovator from the market, sacrificing all of the investment, or could introduce delays allowing the initial data gatherer to enter and compete. An ex-post penalty would also remove the innovator's right to bargain the price down.

4.7.3 Strong property rights will induce CROs to form

The general argument is that intellectual property suffers from high enforcement (monitoring) and valuation costs (Merges 1996; Perritt 1996). In addition, where there are high coordination costs due to many buyers and sellers, Collective Rights Organizations (CROs) will form to minimize transactions costs by centralizing the activities and pooling the rights (Merges 1996).

The commercial market for data reuse and redistribution deviates from the pattern for CRO formation in at least two respects. First, we saw earlier that second comers may compete directly with the initial rights holders raising the potential for strategic bargaining. Merges notes that in markets where the relevant incentives exist, strategic bargaining may impede CROs from emerging.

Second, it is not clear that the commercial market for data reuse and redistribution would support the critical mass of customers and sellers necessary to induce CRO formation (Ginsburg 1997). As noted earlier, the market for databases is generally characterized by niches (NRC 1997a; Reichman and Samuelson 1997). Genome database consumers tend not to purchase financial data. Even the largest market, that of financial data, reflects individual financial metrics and instruments (Tyson and Sherry 1997).

4.7.4 Courts are poor at valuation

Merges (1996) outlines the arguments for why some consider Courts to be inferior to markets at valuation. Like legislatures, courts are vulnerable to lobbying, their proceedings are often reduced to debates between armies of hired experts for opposing parties, and establish precedents that are difficult to overturn.

While Courts may be inferior to markets, our expectation is that, in the long run, given the opportunity to choose otherwise, transacting parties will not resort to Courts too often. As argued earlier, we hope that misappropriation will provide parties with an adequate incentive to bargain.

When the Courts are called upon to value, in misappropriation cases, the liability right ensures a market to assist the Courts in ex post valuation. By allowing the innovation to see the light of day, liability rules use the market to help determine whether the new product competes directly with the initial data gatherer and whether there is a significant market at all.

4.7.5 CROs are better at valuing

In general, Merges argues that CRO (Collective Rights Organization) pricing is determined by professionals engaged in the relevant industry rather than lay judges or legislators and is therefore more likely to accurately infer valuation in the absence of a market. However, his example of ASCAP price-setting is dominated by professionals from the supply-side. ASCAP is an institution by and for producers and artists. The very existence of BMI, a parallel CRO created by broadcasters, challenges the inherent superiority of CRO pricing absent competition.

Moreover, it is not clear that CROs are a better alternate valuation mechanism for the specific problem of commercial data reuse and redistribution. We accept that CROs can reduce coordination costs in markets where there are repeated transactions between many buyers and many sellers. However, as established earlier, in a differentiated market like commercial data reuse and redistribution where users of genome databases are unlikely to purchase real-time financial data, the benefits of reduced coordination costs appear less meaningful. At the same time, CROs that are dominated by one party, as ASCAP is dominated by producers and artists, arguably increase the transactions costs associated with incentives to bargain strategically. Note that this dominance led to the formation of BMI as an alternative (Besen, Kirby, and Salop 1992).

4.7.6 Courts are clogged and time consuming

Aside from questions about their effectiveness at valuation, relying upon the Courts to enforce liability rules also faces the very real constraint that Courts are already heavily backlogged and litigation is a time consuming process. The specter of delays due to Court clog could compromise the effectiveness of misappropriation as an incentive to bargain. Without a realistic threat of ex-post litigation, a data integrator has less incentive to bargain a' priori rather than simply (mis)appropriating the data in question. We will then have introduced an unintended consequence. By overcoming the first-comers strategic bargaining, we may increase the transactions cost of enforcement to the point that the threat of litigation is no longer real.

While Courts face undeniably full schedules, we suggest that the disincentives, due to higher enforcement costs (likelihood of litigation decreases significantly due to Court clog), is lower than is otherwise perceived. First, while valuation may prove time consuming, Courts also may issue preliminary injunctions in advance of valuation proceedings. Injunctions are an effective threat for, where applied, they suspend the integrator's market. Second, allowing the integrator's innovation to see the light of day reveals the market for reuse or redistribution as an aide in assessing competitiveness and impacts on the data gatherer's incentives. Finally, early cases can establish precedents to define subsequent bargaining positions in future cases.

4.7.7 Misappropriation hurts innovation in re-use and re-distribution

A misappropriation statute may preclude some second comers from bringing their innovations into the light of day for fear of costly and time-consuming ex-post litigation. Essentially, innovators became afraid to develop follow-on products (Ginsburg 1990; 1997).

First, it is important to note that some limitations on reuse are necessary. "Free riding discourages investments necessary for innovation, with the result that there are no inventions to imitate. Consumers are better off with the benefits of an innovation that a competitor chooses to reinvent than they would be with no innovation at all (Paepke 1987 at 78)." As a consequence, this does suggest that some innovations at the margin will indeed not see the light of day. Second, it is worth noting the emphasis that our proposal takes on balance. The general intuition is to allow reuse without explicit permission where explicit permission is not granted. The goal is to enable innovation without damaging the incentive to produce from free-riding (Gordon 1992b at note 245). Finally, we argue that by focusing on related markets (e.g. not in direct competition), misappropriation will not impede integration and other follow-on information products.

4.7.8 Misappropriation deters innovation in the primary market

While misappropriation, as a policy, aims to promote progress by encouraging follow-on creation, it does so by granting a monopoly in the market set out by the initial producer. A monopoly introduces the danger of impeding innovation in the primary market. The danger is only exacerbated by the lack of an explicit time limit on the duration of the right to make a misappropriation claim.

First, a limited monopoly is not unjustified in some circumstances. As noted earlier, there is a need to provide an incentive to create (or in the case of data, to gather) in the first place (Paepke 1987). Limits on the monopoly, in the case of data, stem in part from follow-on integrative products that not only explore new markets but also more finely differentiate existing markets to capture deadweight loss (Pindyck and Rubinfeld 1992).

Second, recall the data/presentation distinction. The restraint on competition posed by misappropriation is only for second-comers who would reuse in competition. Second-comers with a better way to gather/produce data may compete directly with any first mover. As noted earlier, there is no property right in the data itself.

Third, integrators with a better selection or arrangement in the initial data gatherers market can demonstrate the viability of the innovation in the marketplace and claim a copyright over the presentation. Two markets then emerge. The integrator may bargain for the right to the underlying data and the initial gatherer may bargain for the right to use the innovative selection and presentation. How "use" of the data or of the selection and presentation relates to the performance copyright is a question for further research (Patterson 1992). Should cross licensing fail, there is always the possibility that both parties sell directly to the consumer; the consumer then integrates the different inputs.

Finally, consider that the promotion of progress includes the initial incentive to produce in the first place. Arguably, once an investment is recovered, extended protection is no longer justified. While a fixed term of protection should not invalidate earlier claims, calculating the optimal misappropriation duration is beyond the scope of this thesis.

We began this chapter by establishing underlying objectives for protecting databases. We then introduced two frameworks for constructing and evaluating policies to satisfy our objectives. Next, we leveragde our technology considerations to argue that databases are a unique form of intellectual property. As a consequence, we found that traditional policy measures for meeting the policy objectives, with respect to data, are inadequate. We present misappropriation as a better alternative. We observed in Chapter 2 that ours are not the only arguments in favor of misappropriation. However, our arguments, couched in the underlying principles for data management, offer a new perspective.

5 Conclusion

In this paper, we explore intellectual property policies for addressing the attribution problem space that stems from data integration. While data integration is not new, modern information technologies in general and the World Wide Web in particular have made data integration an everyday phenomenon. Web portals, comparison sites, personalized pages, and other examples of on-line integration exacerbate tensions about data quality, intellectual property, and data organization. To consider different policy perspectives, we divided the attribution problem space into a number of different dimensions. In this last chapter, we discuss our conclusions from the perspective of these dimensions. We begin with a summary of the paper. We then review our contributions and discuss both limitations and opportunities for future work.

5.1 Summary

We separate the attribution problem space along the dimensions of who, what, where, when, why, and how. We want to know *who* takes data, *what* data they take, *where* the data comes from, *when* the data is taken (i.e. cached vs. real-time), *why* or on whose behalf the data is taken, and *how* the content is used (e.g. in direct competition with the original data provider). By considering the dimensions addressed by different policy measures, we can better understand how the initiatives interact.

In the policy analysis, we first define the status quo approach to each of the dimensions in the attribution problem space. In policy terms, the question of *what* integrators may reuse from other data sources is defined by the *fact* versus *creative work* distinction drawn in (Feist v. Rural 1991). Building from *Feist*, our analysis covers legal precedents governing *who* may take data, *why* and *when* they may take data, and *how* that data may be used. The policy analysis concludes with a review of stakeholders and their respective interests.

We end with a policy formulation exercise. Building from the decision in *Feist*, we identify *misappropriation* as a policy suited to address *who*, *why*, *when* and *how*. We construct a policy to manage the attribution problem space from the intellectual property policy framework. Two economic frameworks for evaluating policy success are presented. The first framework is based upon the prisoner's dilemma; the second is based upon transactions cost economics as applied to entitlement theory. Next, we revisit the attribution problem space in light of these economic metrics. Specifically, building from the database foundations in Part one, we argue for the creativity in structured and semistructured collections of facts. Finally, we present misappropriation as a policy alternative that addresses the stakeholder interests from the policy analysis as evaluated by both economic frameworks.

5.2 Contributions

As noted in Chapter 2, the problem of attribution has been addressed from a technology perspective as well as a policy perspective many times before. However, we believe that our multidimensional depiction of the attribution problem space provides a unique framework for policy analysis. First, our characterization allows us to tie together the myriad policy threads

that cover integration. Second, identifying stakeholders proved more complicated than simply naming base data providers, integrators, and end users. Leaning again on the dimensions of the attribution problem space, we define a taxonomy of stakeholders based upon their interests in the questions of *who*, *what*, *where*, *when*, *why*, and *how*.

The fundamental argument of the policy formulation exercise is that databases are actually the product of two distinct products and processes: gathering and creative selections and arrangements. The creativity inherent in database design and creation belies the commonly accepted cost analysis used to justify strong property rights in data. Ours is not the first work on misappropriation. However, ours is the first, to our knowledge, to draw upon the database literature to inform the policy discussion. In the past, policy makers have addressed all works of information in a uniform fashion, weighing the cost of inducing creative works against the public interest in open access. Accordingly, *Feist* drew a line between non-creative facts and works of information. However, even as stakeholders lobby for new database protections, semistructured models to represent and manipulate data on the Web are blurring the traditional facts versus creativity distinctions.

5.3 Limitations and future work

While this thesis has attempted to cover a great deal of ground, it has also made a number of assumptions and left many issues un-addressed. In this final section, we consider opportunities for future work.

First, this work would benefit from empirical results to reinforce the taxonomy of different stakeholders and integration types. Second, beyond the current review of the policy landscape, we should consider the interaction effects of a host of other policies. The federal government, for example, has policies regarding data documentation that could affect the attribution policy space. Data privacy and security concerns are also an issue. In some circumstances, views and aggregations are deliberately used to anonymize or provide access controls on data (NRC 1997; Ullman and Widom 1997). Third, there are opportunities for an international, comparative analysis of data protections. In this work, we merely raised the European Database Directive as a reference point. However, given the borderless property of the Internet, there is cause for a broader perspective. It would also be interesting to consider whether the same attribution problem space definition is equally applicable to the global perspective.

Finally, in our policy formulation exercise, we need a better economic model of the database industry. No such model currently exists (Reichman and Samuelson 1997; Tyson and Sherry 1997). Current policy draws a distinction between facts and creative works, but applies the same economic models for their creation and distribution. We have offered preliminary arguments for why, from a cost perspective, the distinction is far less obvious. Along a slightly different thread, even as policy-makers have argued for a distinction between facts and creative works, economic models on the value of information treat all data the same. We might like to speculate on whether models on the value of information are useful in articulating a different economic rationale for the (absence of a) distinction between facts and creations.

References

2000. Frequently Asked Questions. bookfinder.com, http://www.bookfinder.com/help/faq/.

Aber, Robert E. 1998. H.R. 2652 Testimony on behalf of Information Industry Association. Before *Subcommittee on Courts and Intellectual Property of the Committee on the Judiciary*, US House of Representatives, Washington, DC. 12 February. http://www.house.gov/judiciary/41143.htm.

Abiteboul, S. 1997. Querying semistructured data. *International Conference on Database Theory (ICDT '97)*, 8-10 January, in Delphi, Greece.

Abiteboul, Serge, Peter Buneman, and Dan Suciu. 2000. *Data on the Web: From Relations to Semistructured Data and XML*. San Francisco, CA: Morgan Kaufmann Publishers.

Abiteboul, Serge, Ricard Hull, and Victor Vianu. 1995. *Foundations of Databases*. Menlo Park: Addison-Wesley Publishing Company.

Abiteboul, S., D. Quass, J. McHugh, J. Widom, and J. Wiener. 1997. The Lorel query language for semistructured data. *International Journal on Digital Libraries* 1 (1):68-88, April.

Abiteboul, S., and V. Vianu. 1997. Querying the Web. *International Conference on Database Theory (ICDT '97)*, 8-10 January, in Delphi, Greece.

Akamai, White Paper. 2001. Turbo-Charging Dynamic Web Sites with Akamai EdgeSuite. Akamai Technologies, Inc., AKAMWP-TCD1201, http://www.akamai.com/en/resources/pdf/Turbocharging_WP.pdf.

Anderson, William C. 1893. *A Dictionary of Law 1893: A Dictionary and Compendium of American and English Jurisprudence*. Ecclesiastic Commonwealth Community, 2 November 2001 [cited 26 January 2002]. http://ecclesia.org/lawgiver/C.asp.

ASCAP. 2001. *About ASCAP: What Is ASCAP* [cited 20 August 2001 2001]. http://www.ascap.com/about/whatis.html.

Bailey, Joseph P. 1998. Intermediation and electronic markets: Aggregation and pricing in Internet commerce. PhD, Technology, Management and Policy, Massachusetts Institute of Technology, Cambridge.

Baird, Douglas G. 1983. Common Law Intellectual Property and the Legacy of International News Service v. Associated Press. *University of Chicago Law Review* 50:411, Spring.

Band, Jonathan. 1998. *The Digital Millennium Copyright Act, analysis* [Web]. Morrision & Foerster, LLP, Washington, D.C., 20 October 1998 [cited June 2000 2000]. http://www.arl.org/info/frn/copy/band.html.

Band, Jonathan. 1998. Testimony on behalf of the Online Banking Association. Before *Subcommittee on Courts, Intellectual Property and the Administration of Justice*, U.S. House of Representatives. 12 February 1998. http://www.house.gov/judiciary/41148.htm.

Band, Jonathan, and Jonathan S. Gowdy. 1997. Sui generis database protection: has its time come? *D-Lib Magazine*, June, http://www.dlib.org/dlib/june97/06band.html.

Bang, Grace. 1997. European Union Protection of Databases: An Overview of the Database Directive. SUNY Buffalo, http://wings.buffalo.edu/Complaw/CompLawPapers/bang.htm.

Baumol, William, and J. Gregory Sidak. 1994. *Toward competition in local telephony*. AEI studies in telecommunications deregulation, *AEI studies in telecommunications deregulation*. Washington, D.C.: American Enterprise Institute for Public Policy Research.

Berkman, H. 1999. Congress Tackles Database Law. *The National Law Journal*, 22 July.

Bernstein, Philip A., and Thomas Bergstraesser. 1999. Meta-data support for data transformations using Microsoft Repository. *IEEE Data Engineering* 22 (1):9-14.

Besen, Stanley M., Sheila N. Kirby, and Steven C. Salop. 1992. An Economic Analysis of Copyright Collectives. *Virginia Law Review* 78:383, February.

Bloomberg, Michael. 1996. *Michael Bloomberg on WIPO database treaty* [Web news posting] [cited 30 November 2000 1996]. http://www.ainfos.ca/A-Infos96/8/0270.html.

Bloomberg News, Staff. 2001. Bidder's Edge Settle Suits on Web Access. *Los Angeles Times*, 2 March, Sec C, p 2.

BMI. 2001. *BMI Backgrounder* [cited 20 August 2001 2001]. http://www.bmi.com/about/backgrounder.asp.

Bohlen, Michael H., Richard T. Snodgrass, and Michael D. Soo. 1996. Coalescing in Temporal Databases. *Twenty-second International Conference on Very Large Data Bases*, 3-6 September, in Bombay, India, pp 180-91.

Bond, Robert. 1996. *European Union Database Law and the Information Society*. Hobson Audley Hopkins & Wood, 1996 [cited 2 July 2000]. http://ds.dial.pipex.com/town/close/gbb67/itlaw/databas.htm.

Borzo, Jeanette. 2001. Searching: Out of order? *Wall Street Journal*, 24 September, Sec E-Commerce (A Special Report), p R13.

Bray, Tim, Jean Paoli, and C. M. Sperberg-McQueen. 1997. Extensible Markup Language (XML). *XML Journal* 2 (4), Fall.

Bressan, S, C Goh, N Levina, A Shah, S Madnick, and M Siegel. 2000. Context Knowledge Representation and Reasoning in the Context Interchange System. *International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies* 12 (2):165-79, September.

Brown, Mark. 2000. *Zagat Survey: 2000/2001 Philadelphia Restaurants* Edited by M. Klein and N. Gottlieb. New York, NY: Zagat Survey, LLC.

Bulfinch, Thomas. 2001. *Bulfinch's Mythology*. Fisher, Bob, April 2001 [cited 26 January 2002 2002]. http://www.webcom.com/shownet/bulfinch/fables/bull20.html.

Buneman, Peter. 1997. Semistructured data. *Sixteenth ACM Symposium on Principles of Database Systems (PODS)*, 13-15 May, in Tucson, AZ.

Buneman, Peter. 2001. Deep linking (unpublished). University of Pennsylvania.

Buneman, P., S. Davidson, M. Fernandez, and D. Suciu. 1997. Adding structure to unstructured data. *International Conference on Database Theory (ICDT `97)*, 8-10 January, in Delphi, Greece.

Buneman, Peter, Alin Deutsch, and Wang-Chiew Tan. 1998. A deterministic model for semistructured data. *Workshop on Query Processing for Semistructured Data and Non-Standard Data Formats*, http://db.cis.upenn.edu/DL/icdt.ps.gz.

Buneman, Peter, Sanjeev Khanna, and Wang-Chiew Tan. 2000. Data Provenance: Some Basic Issues. *Foundations of Software Technology and Theoretical Computer Science*, 13-15 December, in New Delhi, India.

Buneman, Peter, Sanjeev Khanna, and Wang-Chiew Tan. 2001. Why and Where: A Characterization of Data Provenance. *International Conference on Database Theory (ICDT `01)*, 4-6 January, in London, England, http://db.cis.upenn.edu/DL/whywhere.ps.

Buneman, Peter, Keishi Tajima, and Wang-Chiew Tan. 2001. Deep Citation and Efficient Archiving in Digital Libraries. University of Pennsylvania for Digital Libraries Initiatives II Meeting, http://db.cis.upenn.edu/DL/DL-roanoke.pdf.

CADP, Coalition Against Database Piracy. 2000. *H.R. 354: A Balanced Approach* [cited 2 July 2000]. http://www.gooddata.org/quotes.htm.

Calabresi, Guido, and A. Douglas Melamed. 1972. Property Rules, Liability Rules, and

Inalienability: One View of the Cathedral. *Harvard Law Review* 85 (6):1089, April, http://heinonline.org.

Casey, Tim. 1998. H.R. 2652 Testimony on behalf of the Information Technology Association of America. Before *Subcommittee on Courts and Intellectual Property of the Committee on the Judiciary*, US House of Representatives, Washington, DC. 12 February. http://www.house.gov/judiciary/41143.htm.

Chakrabarti, Soumen, Byron Dom, Prabhakar Raghavan, Sridhar Rajagopalan, David Gibson, and Jon Kleinberg. 1998. Automatic resource compilation by analyzing hyperlink structure and associated text. *Proc. 7th International World Wide Web Conference*, 14-18 April 1998, in Brisbane, Australia, http://decweb.ethz.ch/WWW7/1898/com1898.htm.

Chamberlin, Don, James Clark, Daniela Florescu, Jonathan Robie, Jerome Simeon, and Mugur Stefanescu. 2001. *XQuery 1.0. An XML Query Language.* World Wide Web Consortium, 20 December 2001 [cited 7 July 2001 2001]. http://www.w3.org/TR/2001/WD-xquery-20010607/.

Chamberlin, Don, Peter Fankhauser, Massimo Marchiori, and Jonathan Robie. 2001. *XML Query Use Cases: W3C Working Draft 08 June 2001.* World Wide Web Consortium, 20 December 2001 [cited 17 August 2001]. http://www.w3.org/TR/xmlquery-use-cases.

Chawathe, S., H. Garca-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. 1994. The TSIMMIS project: Integration of heterogeneous information sources. *Information Processing Scoiety of Japan*, October, in Tokyo, Japan.

Chawathe, Sudarshan S., Serge Abiteboul, and Jennifer Widom. 1999. Managing Historical Semistructured Data. *Theory and Practice of Object Systems* 24 (4):1, 1999.

Clark, James, and Steve DeRose. 2001. *XML Path Language (XPath) Version 1.0: W3C Recommendation 16 November 1999* [cited 17 August 2001]. http://www.w3c.org/TR/1999/REC-xpath-19991116.

Coase, Ronald. 1988. The Nature of the Firm (1937). In *The Firm, the Market, and the Law.* Chicago, IL: University of Chicago Press.

Cohen, Julie E. 1997. Some reflections on copyright management systems and laws designed to protect them. *Berkeley Technology Law Journal* 12 (1), http://www.law.berkeley.edu/journals/btlj/articles/12_1/Cohen/html/text.html.

Constant, Beth A. 2000. Chalk Talk: The Fair Use Doctrine: Just What Is Fair? *Journal of Law and Education* 29:385, July.

Corlin, Richard F. 1998. H.R. 2652 Statement of the American Medical Association. Before *Subcommittee on Courts and Intellectual Property of the Committee on the Judiciary*, US House of Representatives, Washington, DC. 12 February. http://www.house.gov/judiciary/41146.htm.

Cronk, Denis R. 2000. *Tighter Protection Against Piracy of Online Data: Top NAR Legislative Priority*. National Association of Realtors, 6 January 2000 [cited 30 November 2000]. http://nar.realtor.com/news/2000Releases/January/6.htm.

Cui, Claire Yingwei, and Jennifer Widom. 2000. Practical Lineage Tracing in Data Warehouses. *International Conference on Data Engineering*, February, in San Diego, California, http://www-db.stanford.edu/pub/papers/trace.ps.

Cui, Claire Yingwei, and Jennifer Widom. 2001. Lineage Tracing for General Data Warehouse Transformations. *27th International Conference on Very Large Data Bases (VLDB)*, 11-14 September, in Rome, Italy, http://dbpubs.stanford.edu:8090/pub/2001-5.

Cui, Claire Yingwei, Jennifer Widom, and Janet L. Wiener. 1997 (revised 1999). Tracing the Lineage of View Data in a Datawarehousing Environment. Stanford University, http://www-db.stanford.edu/pub/papers/lineage-full.ps.

databasedata.org. 1999. A Basic Guide to Database Legislation in the 106th Congress. databasedata.org, http://www.databasedata.org/db101/db101.html.

databasedata.org. 1999. Side-By-Side Comparison of Database Protection Bills. databasedata.org, http://www.databasedata.org/DBside-by-side.

deBakker, Bas, and Irsan Widarto. 2001. *An Introduction to XQuery*. X-Hive Corporation, 13 December [cited 13 December 2001]. http://www.perfectxml.com/articles/xml/xquery.asp.

Desai, B. C., P. Goya, and F. Sadri. 1987. Non-first normal form universal relations: an application to information retrieval systems. *Information Systems* 12 (1):49-55, 1987.

Dey, Debabrata, Terence Barron, M., and Veda C. Storey. 1996. A complete temporal relational algebra. *VLDB Journal* 5:167-180.

Dey, Debabrata, and Sumit Sarkar. 1996. A Probabilistic Relational Model and Algebra. *ACM Transactions on Database Systems* 21 (3):339-369, September.

Djavaherian, David. 1998. Hot News and No Cold Facts: NBA v. Motorola and the Protection of Database Contents. *Richmond Journal of Law and Technology* 5 (2), Winter, http://www.richmond.edu/~jolt/v5i2/djava.html.

DOS, Department of State Bureau of Administration. 2001. *Key Officers List, Japan*. U.S. Department of State, 18 October 2001 [cited 2001]. http://www.foia.state.gov/mms/KOH/keypostdetails.asp?post=0&letter=J&id=75.

Drahos, Peter. 1996. *A philosophy of intellectual property*. Brookfield, Vermont: Dartmouth Publishing Company.

Duncan, Daniel C. 1999. H.R. 354 Testimony on behalf of the Software and Information Industry Association. Before *Subcommittee on Courts and Intellectual Property of the Committee on the Judiciary*, US House of Representatives, Washington, DC. 18 March 1999. http://www.house.gov/judiciary/106-dunc.htm.

Duschka, Oliver, and Michael Genesereth. 1997. Answering recursive queries using views. *Sixteenth ACM Symposium on Principles of Database Systems (PODS)*, 13-15 May, in Tucson, AZ.

Duschka, Oliver M. , and Michael R. Genesereth. 1997. Query Planning in Infomaster. *ACM Symposium on Applied Computing*, February, in San Jose, CA.

eBay. 2000. *eBay, Inc.v. Bidder's Edge Inc.*, U.S. District Court for the Northern District of California:LEXIS 13326 (21 July).

Effross, Walter A. 1998. Withdrawl of the reference: rights, rules, and remedies for unwelcomed Web-linking. *South Carolina Law Review* 49:651-593, http://www.wcl.american.edu/pub/faculty/effross/withdrawl.html.

Elgison, Martin, and James M. Jordan. 1997. Trademark cases arise from meta-tags, frames: disputes involve search-engine indexes, web sites within web sites, as well as hyperlinking. *National Law Journal*, 20 October, http://cyber.law.harvard.edu/metaschool/fisher/linking/framing/mixed1.html.

Elkin-Koren, Niva. 1997. Copyright policy and the limits of freedom of contract. *Berkeley Technology Law Journal* 12 (1), http://www.law.berkeley.edu/journals/btlj/articles/12-1/koren.html.

Feist v. Rural. 1991. *Feist Publications, Inc. v. Rural Telephone Service*, U. S. Supreme Court 499:340 (1991).

Ferber, Don. 1991. Tracking Tiger: The use, verifcation, and updating of tiger data. *GIS/LIS*, 1991, in Atlanta, Georgia, pp 230-239.

Ferber, Don. 1992. GIS project documentation: The Wisconsin TIGER Project example. *GIS/LIS (1992)*, 10-12 November, in San Jose, California, pp 221-230.

Fernandez, M., D. Florescu, J. Kang, A. Levy, and D. Suciu. 1997. Strudel: A web site management system. *ACM SIGMOD Conference on Management of Data*, 13-15 May, in Tucson, AZ.

Fernandez, M., D. Florescu, A. Levy, and D. Suciu. 1997. A query language for a web-site management system. *SIGMOD Record* 26 (3):4-11, September.

Fernandez, Mary, and Jonathan Marsh. 2001. *XQuery 1.0 and XPath 2.0 Data Model: W3C Working Draft 7 June 2001*. World Wide Web Consortium, 20 December [cited 7 July 2001]. http://www.w3.org/TR/2001/WD-query-datamodel-20010607/.

Fernandez, Mary, and Jonathan Robie. 2001. *XML Query Data Model: W3C Working Draft 11 May 2000*. World Wide Web Consortium [cited 7 July 2001]. http://www.w3.org/TR/2000/WD-query-datamodel-20000511.

Ferri, Lisa M., and Robert G. Gibbons. 2000. Forgive Us Our Virtual Trespasses: The 'eBay' Ruling. *New York Law Journal*:1, 27 June 2000.

Firat, A., S. Madnick, and M. Siegel. 2000. The Cameleon Web Wrapper Engine. *VLDB Workshop on Technologies for E-Services*, in Cairo, Egypt.

Florescu, Daniela, Alon Y. Levy, and Alberto Mendelzon. 1998. Databse Techniques for the World-Wide-Web: A Survey. *SIGMOD Record* 1998.

Fry, Jason. 2001. Why Shopper's Loyalty To Familiar Web Sites Isn't So Crazy After All. *Wall Street Journal*, 13 August, Sec Marketplace, p B1.

Fujita, Anne K. 1996. The Great Internet Panic: How Digitization is Deforming Copyright Law. *Journal of Technology Law & Policy* 2 (1), http://journal.law.ufl.edu/~techlaw/2/fall96index.html.

Gale Research, Inc. 1999. *Gale Directory of Databases*. Detroit, MI: Gale Research, Inc.

Garland, Susan. 1999. Whose Info Is It Anyway? *Business Week*, 13 September, 114.

Gibbons, Robert. 1992. *Game Theory for Applied Economists*. Princeton, NJ: Princeton University Press.

Ginsburg, Jane C. 1990. Creation and Commercial Value: Copyright Protection of Works of Information. *Columbia Law Review* 90:1865, November.

Ginsburg, Jane C. 1992. No "Sweat"? Copyright and Other Protection of Works of Information

After Feist v. Rural Telephone. *Columbia Law Review* 92:338, March.

Ginsburg, Jane C. 1997. Statement on H.R. 2652: The Collections of Information Antipiracy Act. Before *Subcommittee on Courts, Intellectual Property and the Administration of Justice*, U.S. House of Representatives. 28 October 1997. http://www.house.gov/judiciary/41147.htm.

Goh, Cheng Hian. 1997. Representing and reasoning about semantic conflicts in heterogeneous information systems. Doctor of Philosophy, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Goh, Cheng Hian, S Bressan, S Madnick, and M Siegel. 1999. Context Interchange: New Features and Formalisms for the Intelligent Integration of Information. *ACM Transactions on Office Information Systems*, July.

Goldstein, Paul. 1994. Toward a Third Intellectual Property Paradigm: Comments: Comments on a Manifesto Concerning the Legal Protection of Computer Programs. *Columbia Law Review* 94:2573, December.

Gordon, Wendy J. 1992. On Owning Information: Intellectual Property and the Restitutionary Impulse. *Virginia Law Review* 78:149, February.

Gordon, Wendy J. 1992. Asymmetric Market Failure and Prisoner's Dilemma in Intellectual Property. *University of Dayton Law Review* 17:853, Spring.

Gordon, Wendy J. 1994. Toward a Third Intellectual Property Paradigm: Comments: Assertive Modesty: An Economics of Intangibles. *Columbia Law Review* 94:2579, December.

Gorman, Robert A., and Jane C. Ginsburg. 1993. *Copyright for the Nineties*. Fourth ed. Charlottesville, VA: Michie Company.

Grady, Richard K. 1988. Data lineage in land and geographic information systems (LIS/GIS). *GIS/LIS (88)*, 30 November - 2 December, in San Antonio, Texas.

Green, Robert. 2000. *eBay Revisited.* the Synthesis, 1 July [cited 3 December 2000]. http://www.synthesis.net/columns/websight/07/01.

Grimm, Brothers. 2000. *Grimm's Fairy Tales "Hansel and Gretel"* [Web]. Mordent Software [cited 30 June 2000 2000]. http://www.mordent.com/folktales/grimms/hng/hng.html.

Grimm, Brothers, Josef Scharl Scharl, Jacob Ludwig Carl Grimm, and Wilhelm Grimm. 1976. *The Complete Grimm's Fairy Tales (Pantheon Fairy Tale and Folklore Library)* Edited by

J. Stern: Random House.

Grosso, Paul, and Norman Walsh. 2000. XSL Concepts and Practical Use. *XML Europe 2000*, 12 June, in Paris, France, http://www.nwalsh.com/docs/tutorials/xsl/xsl/slides.html.

Guelich, Scott, Shishir Gundavaram, and Gunther Birznieks. 2000. *CGI Programming with Perl*. 2nd ed. Sebastopol, CA: O'Reilly & Associates, Inc.

H.R. 354. 1999. *Collections of Information Antipiracy Act*. R. H. Coble:To amend title 17, United States Code, to provide protection for certain collections of information., U.S. House of Representatives, 106th Congress, 19 January.

H.R.1858. 1999. *Consumer and Investor Access to Information Act of 1999*. R. T. Bliley:To promote electronic commerce through improved access for consumers to electronic databases, including securities market information databases., U.S. House of Representatives, 106th Congress, 19 May 1999.

H.R. 2652. 1997. *Collections of Information Antipiracy Act*. R. H. Coble:To amend title 17, United States Code, to prevent the misappropriation of collections of information., U.S. House of Representatives, 105th Congress, 9 October.

H.R. 3531. 1996. *Database Investment and Intellectual Property Antipiracy Act of 1996*. R. C. J. Moorehead:To amend title 15, United States Code, to promote investment and prevent intellectual property piracy with respect to databases., U.S. House of Representatives, 104th Congress, 23 May.

Hammack, William. 1998. H.R. 2652 Testimony on behalf of the Association of Directory Publishers. Before *Subcommittee on Courts and Intellectual Property of the Committee on the Judiciary*, US House of Representatives, Washington, DC. 12 February. http://www.house.gov/judiciary/41146.htm.

Hardy, Trotter. 1995. Contracts, Copyright, and Premeption in a Digital World. *Richmond Journal of Law and Technology* 1 (2), http://www.urich.edu/olt/v1i1/hardy.html.

Hardy, Trotter. 1996. Property (and Copyright) in Cyberspace. *The University of Chicago Legal Forum*:217.

Hawkins, Jennifer L. 1997. ProCD, Inc. v. Zeidenberg: Enforceability of shrinkwrap licenses under the Copyright Act. *Richmond Journal of Law and Technology* 3 (1), http://www.richmond.edu/~jolt/v3il/hawkins.html.

Henderson, Lynn O. 1999. H.R. 354 Testimony on behalf of Agricultural Publisher's Association. Before *Subcommittee on Courts and Intellectual Property of the Committee on*

the Judiciary, US House of Representatives, Washington, DC. 18 March 1999. http://www.house.gov/judiciary/106-hend.htm.

Hitchcock, Steve, L. Carr, S. Harris, J. Hey, and W. Hall. 1997. Citation linking: Improving access to online journals. *2nd ACM International Conference on Digital Libraries*, 23-26 July, in Philadelphia, PA, pp 115-122, http://journals.ecs.soton.ac.uk/acmdl97.htm.

Horbaczewski, Henry. 1999. On behalf of the Coalition Against Database Piracy on H.R. 1858, the Consumer and Investor Accessto Information Act of 1999. Before *Subcommittee on Telecommunications, Trade and Consumer Protection of the House Commerce Committee*, US House of Representatives, Washington, DC. 15 June 1999. http://www.gooddata.org/Horbaczewski_testimony.htm.

hotelguide.com. 2001. *Hotelguide.com - Book your accomodation online from our International Hotel Directory*. hotelguide.com [cited 20 August 2001]. http://www.hotelguide.com.

Howe, Dennis, ed. 2000. *Free On-line Dictionary of Computing*: Imperial College Department of Computing.

Hu, Jim. 2000. *MP3.com pays $53.4 million to end copyright suit*. CNET News.com, 15 November, 11:20 am PT [cited 3 December 2000]. http://news.cnet.com/news/0-1005-202-3681102.html.

Huang, Kuan-Tsae, Yang W. Lee, and Richard Y. Wang. 1999. *Quality Information and Knowledge*. Upper Saddle River, NJ: Prentice Hall PTR.

Hunsucker, G.M. 1997. The European Database Directive: Regional stepping stone to an international model? *Fordham Intellectual Property, Media and Entertainment Law Journal* 7.

IFLA, International Federation of Library Associations. 2002. *Committee on Copyright and other Legal Matters*. IFLA, 22 November 2001 [cited September 2001]. http://www.ifla.org/III/clm/copyr.htm.

INS v. AP. 1918. *International News Service v. Associated Press*, U.S. Supreme Court 248:215 (1918).

Japan Youth Hostels, Inc. 2001. *Tokyo, Japan Youth Hostels*. Hostelling International [cited 20 August 2001]. http://www.jyh.or.jp/olhb/JYH-English/jyh.kantou/jyh-7.13.html.

Junnarkar, Sandeep. 1999. Ticketmaster Online-CitySearch buys Sidewalk. *CNET News.com*, 19 July 1999, 12:20 PT, http://www.canada.cnet.com/news/0-1005-200-345004.html.

Kaplan, Carl S. 1999. A search site for search sites is accused of trespassing. *New York Times*, 24 September 1999, http://www.nytimes.com/library/tech/99/09/cyber/cyberlaw/24law.html.

Kaplan, Carl S. 2000. Judge says a spider is trespassing on eBay. *New York Times*, 26 May, http://www.nytimes.com/library/tech/00/05/cyber/cyberlaw/26law.html.

Karjala, Dennis S. 1994. Toward a Third Intellectual Property Paradigm: Comments: Misappropriation as a Third Intellectual Property Paradigm. *Columbia Law Review* 94:2594, December.

Katz, Howard. 2001. *An introduction to XQuery*. IBM developer works XML zone articles, June [cited 13 December 2001]. http://www-106.ibm.com/developerworks/xml/library/x-xqury.html.

Kinko's. 1991. *Basic Books, Inc., Harper & Row Publishers, Inc., John Wiley & Sons, Inc., McGraw-Hill, Inc., Penguin Books USA, Inc, Prentice-Hall, Inc., Richard D. Irwin, Inc., and William Morrow & Co., Inc., v. Kinko's Graphics Corporation*, United States District Court for the Southern District of New York 758:1522 (28 March).

Kirkman, Catherine Sansum. 1998. *Legal Protection of Online Databases*. WebTechniques [cited 2 July 2000]. http://www.webtechniques.com/archives/1998/01/just/.

Kleinberg, Jon. 1998. Authoritative sources in a hyper-linked environment. *Proceedings, 9th ACM-SIAM Symposium on Discrete Algorithms*, http://www.cs.cornell.edu/home/kleinber/auth.pdf.

Klug, A. 1988. On Conjunctive Queries Containing Inequalities. *Journal of the Association for Computing Machinery* 35 (1):146-160.

Konopnicki, D., and O. Shmueli. 1995. W3QS: A query system for the World Wide Web. *21st International Conference on Very Large Data Bases (VLDB)*, 11-15 September, in Zurich, Switzerland, pp 54-65.

Kravitz, Mark. 2001. *$18 and Under: The Guide to Reasonable Dining and Entertainment*. Third ed. Philadelphia, PA: Spirit of `76 Publishing.

Krebs, Brian. 2000. Law profs oppose Court's ban on eBay spidering. *eMarketer*, 3 December, http://www.emarketer.com/enews/20000719_spidering.html.

Krummenacker, Markus. 1995. *Are "Intellectual Property Rights" Justified?* [Web] [cited 11 July 2001].

Kuester, Jeffrey R., and Peter A. Nieves. 1997. What's all the hype about hyperlinking? Thomas, Kayden, Horstemeyer & Risley, L.L.P., http://www.tkhr.com/articles/hyper.html.

Langin, Dan, and James Cary Howell. 2000. ISP Risk Management. *Boardwatch Magazine*, August, 82-6.

Lanter, David P. 1991. Design of a lineage-based meta-data base for GIS. *Cartography and Geographic Information Systems* 18 (4):255-261.

Lanter, David P., and Chris Surbey. 1994. Metadata analysis of GIS data processing, a case study. *International Symposium on Spatial Data Handling (6th)*, 1994, in Edinburgh, Scotland, pp 314-324.

Lasswell, Harold. 1948. The Structure and Function of Communication in Society. In *The Communication of Ideas, A Series of Addresses*, edited by L. Bryson. New York, NY: Institute for Religious and Social Studies, distributed by Harper.

Lederberg, Joshua. 1999. H.R. 354 Testimony on behalf of the National Academy of Sciences, National Academy of Engineering, Institute of Medicine, and the American Association for the Advancement of Science. Before *Subcommittee on Courts and Intellectual Property of the Committee on the Judiciary*, US House of Representatives, Washington, DC. 18 March 1999. http://www.house.gov/judiciary/106-pinc.htm.

Lee, T., and S. Bressan. 1997. Multimodal Integration of Disparate Information Sources with Attribution. *ER 97 Workshop on Information Retrieval and Conceptual Modeling*, November, in Los Angeles, CA.

Lee, T., S. Bressan, and S. Madnick. 1997. Source Attribution for Querying Against Semi-structured Documents. MIT Sloan School of Management, Sloan WP#4042 CISL WP#99-01.

Lee, T., S. Bressan, and S. Madnick. 1998. Source Attribution for Querying Against Semi-stuctured Documents. *Workshop on Web Information and Data Management, Seventh International ACM Conference on Information and Knowledge Management*, 3-7 November, in Bethesda, MD.

Lee, T., M. Chams, R. Nado, S. Madnick, and M. Siegel. 1999. Information Integration with Attribution Support for Corporate Profiles. *Eighth International ACM Conference on Information and Knowledge Management (CIKM)*, 2-6 November, in Kansas City, KS, pp 423-430.

Lenz, Evan. 2001. *XQuery: Reinventing the Wheel?* XYZFind Corp. [cited 13 December

2001]. http://xmlportfolio.com/xquery.html.

Let's Go, Inc. 1993. *Let's Go: Germany, Austria & Switzerland* Edited by G. W. Rodkey. New York, NY: St. Martin's Press.

Levy, Alon Y. 2000. Logic-Based Techniques in Data Integration. In *Logic Based Artificial Intelligence*, edited by J. Minker: Kluwer Publishers.

Levy, Alon Y., Anand Rajaraman, and Joann J. Ordille. 1996. Querying Heterogeneous Information Sources Using Source Descriptions. *22nd International Conference on Very Large Data Bases (VLDB)*, 3-6 September, in Bombay, India.

Lindemans, Micha F. 2000. *The Encyclopedia Mythica* [cited 6/30/2000 2000]. http://www.pantheon.org/mythica/areas/greek.

Linn, Anne. 2000. *History of Database Protection: Legal Issues of Concern to the Scientific Community*. National Research Council, 3 March 2000 [cited 2 July 2000]. http://www.codata.org/codata/data_access/linn.html.

Litman, Jessica. 1992. After Feist. *University of Dayton Law Review* 17.

Liu, H. C., and K Ramamohanarao. 1994. Algebraic equivalences among nested relational expressions. The University of Melbourne, Technical Report 94/4, http://http://www.cs.mu.oz.au/publications/tr_db/mu_94_04.ps.gz.

Liu, Joseph P. 2001. Owning digital copies: Copyright law and the incidents of copy ownership. *William and Mary Law Review* 42:1245-1366, April, 2001.

Lutzker, Arnold P. 1999. *Primer on the Digital Millennium*. Lutzker and Lutzker, LLP, Washington, D.C., 5 February 1999 [cited June 2000]. http://www.arl.org/info/frn/copy/primer.html.

MacMillan, Robert. 2000. Sen. DeWine calls for database bill next year. *Newsbytes*, 26 October, 10:06 AM EST, http:////www.newsbytes.com/news/00/157254.html.

Mahoney, Paul G. 1997. Technology, Property Rights in Information, and Securities Regulation. *Washington University Law Quarterly* 75 (2):815, Summer.

Maier, David. 1983. *The theory of relational databases*. Rockville, Maryland: Computer Science Press.

Marino, Fabio. 2000. *Database Protection in the European Union* [cited 2 July 2000]. http://www.jus.unitn.it/cardozo/Review/Students/Marino1.html.

Markon, Jerry. 2001. E-Business: The Web @ Work/Willkie Farr & Gallagher. *Wall Street Journal*, 30 April, Sec E-Business, p B5.

McDermott, Terry. 1999. H.R. 354 Testimony on behalf of National Association of Realtors. Before *Subcommittee on Courts and Intellectual Property of the Committee on the Judiciary*, US House of Representatives, Washington, DC. 18 March 1999. http://www.house.gov/judiciary/106-mcde.htm.

McHugh, J., S. Abiteboul, R. Goldman, D. Quass, and J. Widom. 1997. Lore: A database management system for semistructured data. *ACM SIGMOD Record* 26 (3):54-66, 1997.

Mendelzon, Alberto, George A. Mihaila, and Tova Milo. 1996. Querying the World Wide Web. *Fourth International Conference on Parallel and Distributed Information Systems (PDIS)*, 18-20 December, in Miami, FL, pp 80-91.

Mendelzon, Alberto, and Tova Milo. 1997. Formal models of Web queries. *Sixteenth ACM Symposium on Principles of Database Systems (PODS)*, 13-15 May, in Tucson, AZ, pp 134-143.

Merges, Robert P. 1994. Toward a Third Intellectual Property Paradigm: Comments: Of Property Rules, Coase, and Intellectual Property. *Columbia Law Review* 94:2655, December.

Merges, Robert P. 1996. Contracting into Liability Rules: Intellectual Property Rights and Collective Rights Organizations. *California Law Review* 84 (5):1293, October, http://www.sims.berkeley.edu/BCLT/pubs/merges/contract.htm.

Merges, Robert P., Peter S. Menell, Mark A. Lemley, and Thomas M. Jorde. 1997. *Intellectual Property in the New Technological Age*. New York: Aspen Law & Business, Aspen Publishers, Inc.

Mihaila, George A., Louiqa Raschid, and Maria Esther Vidal. 1999. Querying "Quality of data" metadata. *IEEE Metadata*, 1999, http://www.computer.org/conferen/proceed/meta/1999/papers/65/gmihaila.html.

Milgrom, Paul, and John Roberts. 1992. *Economics, Organization and Management*. Englewood Cliffs, NJ: Prentice Hall.

Minker, Jack, ed. 1988. *Foundations of Deductive Databases and Logic Programming*. Los Altos: Morgan Kaufmann Publishers, Inc.

Monster.com. 2000. *Monster.com Joins Coalition Against Database Piracy*, 26 September

2000 [cited 30 November 2000]. http://www.gooddata.org/monster.htm.

Motro, Amihai. 1996. Panorama: a database system that annotates its answers to queries with their properties. *Journal of Intelligent Information Systems* 7 (1):51-73.

Motro, Amihai, and Igor Rakov. 1998. Estimating the quality of databases. *Flexible Query Answering Systems. Third International Conference, FQAS'98. Proceedings*, 13-15 May, in Roskilde, Denmark, pp 298-307.

MP3.com. 2000. *UMG Recordings, Inc. v. MP3.com, Inc.*, United States District Court for the Southern District of New York:LEXIS 13293 (6 September).

Nazareth, Annette L. 1999. Prepared statement on behalf of the Securites and Exchange Commission concerning H.R. 1858. Before *Subcommittee on Finance and Hazardous Materials*, U.S. House of Representatives, Washington, D.C. 30 June 1999.

NBA v. Motorola. 1997. *National Basketball Association v. Motorola, Inc.*, 2nd Cricuit 105:841.

NCID, National Center for Infectious Diseases. 2001. Travelers' Health: Health Information for Travelers to East Asia. U.S. Department of Health and Human Services, Centers for Disease Control (CDC), http://www.cdc.gov/travel/eastasia.htm.

Neal, James G. 1999. H.R. 354 Testimony on behalf of American Association of Law Libraries, American Library Association, Association of Research Libraries, Medical Library Association, and Special Libraries Association. Before *Subcommittee on Courts and Intellectual Property of the Committee on the Judiciary*, US House of Representatives, Washington, DC. 18 March. http://www.house.gov/judiciary/106-neal.htm.

Nestorov, S., S. Abietboul, and R. Motwani. 1997. Inferring structure in semistructured data. *Workshop on Management of Semistructured Data in Conjunction with ACM SIGMOD*, 13-15 May, in Tucson, AZ.

NFL v. Delaware. 1977. *National Football League v. State of Delaware*, F. Supp. 435:1372.

Nicolas, Jean-Marie. 1982. Logic for Improving Integrity Checking in Relational Data Bases. *Acta Informatica* 18:227-53.

Nimmer, Raymond T. 1998. Breaking Barriers: The Relation Between Contract and Intellectual Property Law. *Berkeley Technology Law Journal* 13:827, Fall.

Nissen, Dinah, and Jamie Barber. 1996. The EC Database Directive. *In-House Lawyer*, May, http://www.harbottle.co.uk/pubs/may96.htm.

Nottrott, Rudolf W., Matthew B. Jones, and Mark Schildhauer. 1999. Using XML-structured metadata to automate quality assurance processing for ecological data. *IEEE Metadata*, http://www.computer.org/conferen/proceed/meta/1999/papers/64/rnottrott.html.

NRC, National Research Council. 1997. *Bits of Power: Issues in Global Access to Scientific Data*. Computer Science and Telecommunications Board, *Computer Science and Telecommunications Board*. Washington, DC: National Academy Press.

NRC, National Research Council. 1997. *For the Record: Protecting Electronic Health Information*. Computer Science and Telecommunications Board, *Computer Science and Telecommunications Board*. Washington, DC: National Academy Press.

NRC, National Research Council. 1999. *A Question of Balance: Private Rights and Public Interest in Scientific and Technical Databases*. Commission on Physical Sciences, Mathematics, and Applications, *Commission on Physical Sciences, Mathematics, and Applications*. Washington, DC: National Academy Press.

NRC, National Research Council. 2000. *The Digital Dilemma: Intellectual Property in the Information Age*. Engineering and Physical Sciences, *Engineering and Physical Sciences*. Washington, DC: National Academy Press.

Olsen, Stefanie. 1999. *eBay inks deal with auction search site*. CNET News.com, 1 December 1999 2:40 pm PST [cited 3 December 2000]. http://news.cnet.com/news/0-2007-300-1475546.html.

OMM, O'Melveney & Meyers LLP. 1999. *Copyright Law and the Internet*. O'Melveney & Meyers LLP, 19 November 1998 [cited 16 April 1999]. http://www.omm.com/ilpg/ip/copyright.html.

O'Rourke, Maureen A. 1997. Copyright Preemption After the ProCD Case: A Market-Based Approach. *Berkeley Technology Law Journal* 12 (1), http://www.law.berkeley.edu/journals/btlj/articles/12-1/ORourke.html.

O'Rourke, Maureen A. 1999. Progressing Towards a Uniform Commercial Code for Electronic Commerce or Racing Towards Nonuniformity? *Berkeley Technology Law Journal* 14 (2), http://www.law.berkeley.edu/journals/btlj/articles/14_2/O'Rourke/html/reader.html.

OSTP, Office of Science and Technology Policy. 1999. Administration testimony on HR 354. Before *Subcommittee on Courts and Intellectual Property of the Committee on the Judiciary*, US House of Representatives, Washington, DC. 18 March 1999. http://www.whitehouse.gov/WH/EOP/OSTP/html/19993_19_2.html.

Paepke, C. Owen. 1987. An Economic Interpretation of the Misappropration Doctrine: Common Law Protection for Investments in Innovation. *High Technology Law Journal.*

Papakonstantinou, Y., S. Abiteboul, and H. Garca-Molina. 1996. Object fusion in mediator systems. *22nd International Conference on Very Large Data Bases (VLDB)*, 3-6 September, in Bombay, India.

Papakonstantinou, Y., H. Garca-Molina, and J. Widom. 1995. Object exchange across heterogeneous information sources. *International Conference on Data Engineering*, in Taipei, Taiwan, pp 251-260.

Patterson, L. Ray. 1992. Copyright Overextended: A Preliminary Inquiry Into the Need for a Federal Statute of Unfair Competition. *Dayton Law Review* 17:385, Winter.

Perritt, Henry H. Jr. 1996. Property and Innovation in the Global Information Infrastructure. *The University of Chicago Legal Forum*:261.

Peters, Marybeth. 1999. H.R. 354 Testimony for the U.S. Copyright Office. Before *Subcommittee on Courts and Intellectual Property of the Committee on the Judiciary*, US House of Representatives, Washington, DC. 18 March 1999. http://www.house.gov/judiciary/106-pete.htm.

Phelps, Charles E. 1999. H.R. 354 Testimony on behalf of the Association of American Universities, the American Council of Education, and the National Association of State Universities and Land-Grant Colleges. Before *Subcommittee on Courts and Intellectual Property of the Committee on the Judiciary*, US House of Representatives, Washington, DC. 18 March 1999. http://www.house.gov/judiciary/106-phel.htm.

Pincus, Andrew J. 1999. H.R. 354 Testimony for the U. S. Department of Commerce. Before *Subcommittee on Courts and Intellectual Property of the Committee on the Judiciary*, US House of Representatives, Washington, DC. 18 March 1999. http://www.house.gov/judiciary/106-pinc.htm.

Pindyck, Robert S., and Daniel L. Rubinfeld. 1992. *Microeconomics*. Second ed. New York, NY: Macmillan Publishing Co.

Planet, Lonely. 2001. *Worldguide, Destination: Tokyo*. Lonely Planet [cited 20 August 2001]. http://www.lonelyplanet.com/destinations/north_east_asia/tokyo/attractions.htm.

Pollack, Malla. 1999. The Right to Know?: Delimiting Database Protection at the Juncture of the Commerce Clause, the Intellectual Property Clause, and the First Amendment. *Cardozo Arts & Entertainment Law Journal* 17.

Posner, Richard A. 1992. *Economic Analysis of Law*. 4th ed. Boston, MA: Little, Brown.

Princeton v. MDS. 1992. *Princeton University Press, Macmillan, Inc. and St. Martin's Press, Inc. v. Michigan Document Services, Inc. and James M. Smith*, U. S. District Court for the Eastern District of Michigan, Southern Division 1992:13257 (2 April).

Princeton v. MDS. 1996. *Princeton University Press, Macmillan, Inc. and St. Martin's Press, Inc. v. Michigan Document Services, Inc. and James M. Smith*, United States Court of Appeals for the Sixth Circuit 99:1381 (8 November).

ProCD v. Zeidenberg. 1996. *ProCD v. Zeidenberg*, 7th Circuit 908:640.

Quass, D., A. Rajaraman, Y Sagiv, J. Ullman, and J. Widom. 1995. Querying semistructured heterogeneous information. *Fourth International Conferenc on Deductive and Object-Oriented Databases*, in Singapore, pp 436-445.

Quass, D., J. Widom, R. Goldman, K. Haas, Q. Luo, J. McHugh, S. Nestorov, A. Rajaraman, H. Rivero, S. Abiteboul, J. Ullman, and J. Wiener. 1996. LORE: A Lightweight Object REpository for semistructured data. *ACM SIGMOD International Conference on Management of Data*, June, in Montreal, Canada.

Raggett, Dave. 2000. *Adding a touch of style*. W3C, 29 August 2000 [cited 23 October 2001]. http://www.w3.org/MarkUp/Guide/Style.

Raggett, Dave. 2001. *Getting started with HTML*. W3C, 4 June 2001 [cited December 2000]. http://www.w3.org/MarkUp/Guide/.

Ramakrishnan, Raghu, and Johannes Gehrke. 2000. *Database Management Systems*. 2nd ed. Boston, MA: McGraw-Hill.

Raskind, Leo J. 1991. The Misappropriation Doctrine as a Competitive Norm of Intellectual Property Law. *Minnesota Law Review* 75:875, February.

Raul, Alan Charles, Edward R. McNicholas, and Claudia A. von Pervieux. 2000. *Who Owns the Data? Evolving Protections for Facts, Secrets and Personal Information in Cyberspace* [Web]. Washington, D.C. Office of Sidley & Austin, April 2000 [cited 11 December 2000]. http://www.sidley.com/cyberlaw/features/protect.asp.

Reichman, J.H., and Pamela Samuelson. 1997. Intellectual property rights in data? *Vanderbilt Law Review* 50, January.

Reichman, J. H., and Paul F. Uhlir. 1999. Database protection at the crossroads: Recent developments and their impact on science and technology. *Berkeley Technology Law*

Journal 14 (2), http://www.law.berkeley.edu/journals/btlj/articles/14_2/Reichman/html/reader.html.

RIAA. 2000. *MP3.com Lawsuit Q&A*. Recording Industry Association of America [cited 3 December 2000]. http://www.riaa.com/MP3lawsuit.cfm.

Rob, Peter, and Carlos Coronel. 1997. *Databsae Systems: Design, Implementation, and Management*. Cambridge, MA: Course Technology, International Thomson Publishing.

Rosenbaum, David E. 2000. Database Legislation Spurs Fierce Lobbying. *New York Times*, 5 June, Sec A, p 14, http://www.gooddata.org/NYT.htm.

Rosenthal, A., and E. Sciore. 1999. Security administration for federations, warehouses, and other derived data. *IFIP WG11.3 Conference on Database Security*, http://www.cs.bc.edu/~sciore/papers/IFIP99.pdf.

Rosenthal, A., and E. Sciore. 1999. Administering propagated metadata in large, multi-layer database systems. *IEEE Workshop on Knowledge and Data Exchange*, 7 November, http://www.cs.bc.edu/~sciore/papers/KDEX99.pdf.

Roth, Mark A., Henry F. Korth, and Abraham Silberschatz. 1988. Extended Algebra and Calculus for Nested Relational Databases. *ACM Transactions on Database Systems* 13 (4):389-417, December.

Rough Guides, Travel. 2001. *Rough Guide Travel: Tokyo*. Rough Guide Travel [cited 20 August 2001 2001]. http://travel.roughguides.com/content/10072/22912.htm.

S. 95. 1999. *Trading Information Act*. S. J. McCain:To amend the Communications Act of 1934 to ensure that public availability of information concerning stocks traded on an established stock exchange continues to be freely and readily available to the public through all media of mass communication., U.S. Senate, 106th Congress, 1st session, 19 January 1999.

S. 2291. 1998. *Collections of Information Antipiracy Act*. S. R. Grams:A bill to amend title 17, United States Code, to prevent the misappropriation of collections of information, U.S. Senate, 105th Congress, 10 July.

Sableman, Mark. 1999. Link Law: The emerging law of Internet hyperlinks. *Communication Law and Policy* 4 (4):557-601, http://www.ldrc.com/cyber2.html.

Sadri, Fereidoon. 1991. Modeling uncertainty in databases. *International Conference on Data Engineering*, 8-12 April, in Kobe, Japan, pp 122-131.

Sadri, Fereidoon. 1994. Aggregate operations in the information source tracking method. *Theoretical Computer Science* 133 (2):421-442, 24 October.

Sadri, Fereidoon. 1995. Information source tracking method: efficiency issues. *IEEE Transactions on Knowledge and Data Engineering* 7 (6):947-954, December.

Sagiv, Yehoshua, and Mihalis Yannakakis. 1980. Equivalences Among Relational Expresions with the Union and Difference Operators. *Journal of the Association for Computing Machinery* 27 (4):633-655, October.

Samuelson, Pamela. 1992. Copyright Law and Electronic Compilations of Data. *Communications of the ACM* 35 (2), February.

Schek, H. -J., and P. Pistor. 1982. Data Structures for an Integrated Data Base Management and Information Retrieval System. *8th International Conference on Very Large Data Bases*, 8-12 September 1982, in Mexico City, Mexico, pp 197-207.

Schek, H. -J., and M. H. Scholl. 1986. The Relational Model with Relation-Valued Attributes. *Information Systems* 11 (2):137-147.

Scholl, M. H. 1992. Extensions to the relational data model. In *Conceptual modelling, databases, and CASE: An integrated view of information systems development*, edited by L. P. and R. Zicari. New York: Jon Wiley & Sons.

SEC, U.S. Securities and Exchange Commission. 1999. Special Study: On-Line Brokerage: Keeping Apace of Cyberspace. U.S. Securities and Exchange Commission, http://www.sec.gov/news/studies/cyberspace.htm.

Shapiro, Carl, and Hal R. Varian. 1999. *Information Rules: A Strategic Guide to the Network Economy*. Boston, MA: Harvard Business School Press.

Shrager, Heidi J. 2001. E-Business: The Web @ Work/Zagat Survey. *Wall Street Journal*, 20 August 2001, Sec E-Business, p B6.

Sony v. Universal. 1984. *Sony Corp. v. Universal City Studios, Inc.*, U. S. Supreme Court 464:417.

Spaulding, Michelle L. 1998. *The doctrine of misappropriation* [Web]. Harvard Law School, 21 March 1998 [cited December 1999]. http://cyber.law.harvard.edu/metaschool/fisher/linking/doctrine/.

staff. 2000. *Federal judge says MP3.com willfully violated music copyrights*, 6 September 2000, 2:53P EDT [cited 4 January 2001].

http://www.cnn.com/2000/LAW/09/06/mp3.lawsuit.

Tabke, Brett. 1999. *PriceMan Sued by MySimon* [Web]. Saerch Engine World.com, 24 September 1999 [cited 11 December 2000 2000]. http://www.searchengineworld.com/news/lawsuit.htm.

Taylor, Chris, Peter Turner, Joe Cummings, and et al. 1997. *South-East Asia on a shoestring*. Ninth ed. Melbourne, Australia: Lonely Planet Publications.

Terry, Andrew. 1988. Misappropriation of a Competitor's Trade Values. *The Modern Law Review* 51:296, May.

Ticketmaster v. Microsoft. 1997. *Ticketmaster Corp. v. Microsoft Corp.*, 97:3055PP (settled).

Total News. 1997. *Washington Post Company v. Total News Inc.*, S. D. N. Y. 97:1190.

Transradio Press Service. 1937. *Twentieth Century Sporting Club, Inc. v. Transradio Press Service*, New York Supreme Court 300:159.

Tsur, D., J. Ullman, S. Abiteboul, C. Clifton, R. Motwani, S. Nestorov, and A. Rosenthal. 1998. Query flocks: A generalization of association-rule mining. *ACM SIGMOD International Conference on Management of Data*, June, in Seattle, WA, pp 1-12.

Tyson, L, and E Sherry. 1997. Statutory protection for databases: economic and public policy issues. Information Industry Association.

Tzafestas, Elpida. 2000. Toward Adaptive Cooperative Behavior. *Proceedings of the Simulation of Adaptive Behavior Conference*, September, in Paris, France.

U.S. v. Microsoft. 2001. *United States of America v. Microsoft Corporation*, U.S. Court of Appeals for the District of Columbia Circuit 253:34 (28 June).

Ullman, Jeffrey D. 1988. *Principles of database and knowledge-base systems, volume 1*. Principles of Computer Science, Edited by A. V. Aho and J. D. Ullman. 2 vols. Vol. 1, *Principles of Computer Science*. Rockville, Maryland: Computer Science Press.

Ullman, Jeffrey D. 1989. *Principles of database and knowledge-base systems, volume 2*. Principles of Computer Science, Edited by A. V. Aho and J. D. Ullman. 2 vols. Vol. 2, *Principles of Computer Science*. Rockville, Maryland: Computer Science Press.

Ullman, Jeffrey D., and Jennifer Widom. 1997. *A First Course in Database Systems*. New Jersey: Prentice-Hall, Inc.

Van de Sompel, Herbert, and Patrick Hochstenbach. 1999. Reference linking in a hyrid library environment. *D-Lib Magazine* 5 (4), http://www.dlib.org/dlib/april99/van_de_sompel/04vande_sompel-pt1.html.

Van Gelder, Allen, and Rodney Topor. 1991. Safety and Translation of Relational Calculus Queries. *ACM Transactions on Database Systems* 16 (2):235-78, June.

Walsh, N. 1997. Introduction to XML. *XML Journal* 2 (4), Fall.

Wang, Richard, and Stuart Madnick. 1990. A Polygen Model for Heterogeneous Database Systems: The Source Tagging Perspective. *16th International Conference on Very Large Data Bases (VLDB)*, 13-16 August, in Brisbane, Australia.

Warren Publishing v. Microdos. 1997. *Warren Publishing, Inc. v. Microsods Data Corp*, United States Court of Appeals, Eleventh Circuit 93:8474 (10 June).

Wiederhold, G. 1992. Mediators in the architecture of future information systems. *IEEE Computer* 25 (3):38-49, March.

Winokur, Marilyn. 1999. H.R. 354 Testimony on behalf of Thomson Corporation and the Coalition Against Database Piracy. Before *Subcommittee on Courts and Intellectual Property of the Committee on the Judiciary*, US House of Representatives, Washington, DC. 18 March 1999. http://www.house.gov/judiciary/106-wino.htm.

Wolverton, Troy. 2000. *Judge bars Bidder's Edge Web crawler on eBay*. CNET News.com, 25 May 2000, 12:30 PST [cited 3 December 2000]. http://news.cnet.com/news/0-1007-200-1948171.html.

Wong, Stephanie. 1999. Estimated $4.35 billion in ecommerce sales at risk each year. Zona Research, Inc., http://www.zonaresearch.com/info/press/99-jun30.htm.

Woodruff, Allison, and Michael Stonebraker. 1997. Supporting fine-grained data lineage in a database visualization environment. *Proceedings of the 13th International Conference on Data Engineering*, April, in Birmingham, England, pp 91-102.

Zuckerman, Gregory, and Rebecca Buckman. 1999. Data Providers Face Internet Challengers. *Wall Street Journal*, 21 September, p C1.