Attribution Principles for Data Integration: Technology and Policy Perspectives

by
Thomas Y. Lee

S.M. Technology and Policy
Massachusetts Institute of Technology, 1994

A.B. Political Science; B.S. Symbolic Systems
Stanford University, 1992

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
February 2002

Signature of Author: Thomas Y Lee
Technology, Management and Policy, Engineering Systems Division
31 January 2002

Certified by: Stuart E Madnick
Stuart E. Madnick
J.N. Maguire Professor of Information Technology
Thesis Supervisor

Certified by: Lee McKnight
Lee McKnight
Associate Professor of International Communications, Tufts University

Certified by: Peter Szolovits
Peter Szolovits
Professor of Electrical Engineering and Computer Science

Accepted by: Daniel Hastings
Daniel Hastings
Director, Technology and Policy Program
Associate Director, Engineering Systems Division

Attribution Principles for Data Integration: Technology and Policy Perspectives

by
Thomas Y. Lee

Submitted to the Engineering Systems Division
on 31 January 2002
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy
in Technology, Management and Policy

Abstract

This thesis addresses problems of attribution that arise from the data integration that is exemplified by data re-use and re-distribution on the Web. We present two different perspectives. We begin with a simple definition of attribution, asking *what* data are we interested in and *where* does it come from? A formal model and its properties are defined, implementation in an extended relational algebra is described, and application to semistructured data on the Web is discussed. However, because the problem is more than simply *what* and *where*, we then expand the scope of our analysis. 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.

Our formal model of attribution is developed in the established foundation of the Domain Relational Calculus (DRC). Three distinct types of attribution are identified: comprehensive, source, and relevant. For each type, we consider the attribution of equivalent DRC expressions, attribution for composed queries, and granularity. An algebra is presented to implement the model. The extended algebra is closed, reduces to the standard relational algebra, and is a consistent extension of the standard algebra.

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

To Nicholas Patrick Bailey,
who has patiently waited for his godfather to finish this thesis.

ACKNOWLEDGEMENTS

If you thought that writing a doctoral dissertation was difficult, try thanking all of those people who made your work possible without turning your acknowledgements into an autobiography. Imagine, in addition, the irony of attempting to write acknowledgements for a thesis that is about attribution. Given the alternative of a generic note "to the Academy and all who made this possible," I will risk the sin of omission and preemptively ask forgiveness of any who are unintentionally unnamed.

First and foremost, I must thank the members of my thesis committee. For unfathomable reasons, they continued to bear with me throughout this process despite my best efforts to bollix the process. As a digression, allow me to confirm, based upon vast empirical evidence from a sample size of one, that accepting a job and leaving the Institute before finishing your thesis is a bad idea. Personal decisions not withstanding, the document has been accepted for submission. Despite his own commitments, Stuart Madnick adjusted to my schedule whenever I appeared in town. Through him, my thesis drafts have logged more frequent flier miles than I have. His research not only inspired this work but also led me to pursue an academic career in a business school. Beyond the thesis, his advice on curriculum development and case teaching have been invaluable in my first academic position. Lee McKnight took me on as a Master's degree candidate when I first came to MIT. Fortunately, and perhaps foolishly, he was willing to again assume the role of advisor when I chose to return to MIT. His dedication was all the more evidenced in his participation, despite having to commute from Tufts University to do so. Finally, I would like to thank Peter Szolovits. I first consulted with Peter on my Master's degree many years ago. Little did he know that would later entail hiking repeatedly from NE43 to the MIT hinterlands of the Sloan School. Throughout, he has provided invaluable guidance, counseling and support. In particular, I am grateful for his honest evaluation of my work and for the long evening conversations on pedagogy and on working with doctoral students from different academic traditions.

But while my committee provided the foundation, credit must also go to the virtual faculty that assembled over the years, sacrificing their time and energy to tutor me, act as a sounding board, and review drafts. Their number includes Dr. Michael Siegel of the MIT Sloan School of Management, Dr. Arnon Rosenthal of the MITRE Corporation, Dr. Stéphane Bressan of the National University of Singapore, Professor Richard Wang of Boston University, Professor Cheng Hian Goh of the National University of Singapore, Professor Dan Hunter in Legal Studies at the Wharton School, Professor Joseph Liu of the Boston College School of Law, and Professor Joseph Bailey of the University of Maryland. Nancy Lee, Tony Eng, Aykut Firat, Victor Luchangco, Sean McGrew, and Erik Duerr also read portions of this work and provided feedback and encouragement. It goes without saying that, despite the best efforts of all those involved, numerous errors inevitably remain and are the sole responsibility of the author.

In addition to intellectual input, this final document would not have materialized but for the shepherding of a tremendous team of administrative personnel. Sydney Miller has borne the highest cost of keeping up with my intransigence. Yubettys Baez of the Sloan School has offered enthusiastic support and encouragement every time I see her. I am equally thankful

for Jean Marie De Jordy in TPP, Dee-Dow Chase at CAES, and Su Chung at CTPID. Thanks also go to Gail Hickey, Rene Smith, and Agnes Chow, who have all since moved on to other positions but whose presence is still felt.

The enormous investment of human capital was complemented by generous financial support. Through sponsorship of the Context Interchange Research Project, PricewaterhouseCoopers, TRW, the MITRE Corporation, ARPA and the USAF/Rome Laboratory under F30602-93-C0160, and the MIT Lean Sustainment Initiative have all been participants in aspects of this work. The MIT Center for Advanced Educational Services under NSF-NCR-9307548 and ARPA-A0-B231 provided support when I first returned to MIT to pursue doctoral studies.

As alluded to earlier, even as I complete this document, I have already begun an academic appointment at the University of Pennsylvania. Tremendous thanks are owed to Professors Steven Kimbrough and Balaji Padmanabhan for their support and encouragement. Thanks also go to my Department Chairs, Professors Howard Kunreuther and Paul Kleindorfer, for their encouragement and patience. Running with Professor Christian Terwiesch has offered relief. Ann, Kim, Barbara, Cynthia, and Marge, of the department administration, have also been incredibly supportive.

While MIT is unquestionably a stellar academic and research institution, this thesis document, and the education that came with it, is at least as much, if not more, a product of the intellectual community of scholars that embraces MIT and greater Boston. Nowhere is that community better represented than in Ashdown House, my home of eight plus years. One cannot think of Ashdown House without first thinking of Vernon and Beth Ingram, the House Masters, who imbued a building of brick with spirit and life. Ashdown is no less represented by House manager Christine Vardaro, who cares for the building as though it were her own.

While I hung my hat in 404A for more than six years, the community extends far beyond that. There was poker, bridge, Settler's and dim sum ("How could you be so stupid!") There was Rosie's Place, Christmas in April, Ivy's Plus softball, MIT Ultimate, BUDA, Park Street Church, and Café Small Group. And of course the Ashdown House Harriers! To you who have laughed with me and cried with me, suffered with me or simply suffered me, run with me and iced/rehabbed with me, prayed with me or simply for me, you are not only my community but also His body. This thesis (and the miracle it represents) is of your making and His. In the wild year post-MIT but pre-completion, specific thanks for prayers, humor, and encouragement go to Joe and Wendy, Pat and Debbie, Carl and Dianne, the Houlahans, the Nicholsons, Charlene and Sean, Tony, Anita, Lee, Johnny, and Mrs. D. Pat's shared struggles and accountability to Chris and Erik, commiserating brothers who are all running the race and fighting the fight, have proven particularly meaningful.

And in the final months, as things have reached a fever pitch in my Boston-Philadelphia commute, special mention must be made of Pat and Debbie, Tony, and especially Charlene, Sean and Erik. I would arrive on their doorsteps between 11 p.m. and 4 a.m. I have slept in their homes, received rides in their cars, and eaten their food. I have spent as many nights in Erik Duerr's Boston guest room as in my own Philadelphia apartment.

Community, of course, begins at home. I thank my parents, who started me on this path and continue to challenge me through the example of their own lives, always growing. Thank you also to my brother, who provides me with more than just inspiration. He has not only encouraged but participated in my adventures, from higher education to running through the Canadian Rockies. His faith, confidence, and trust remind me of what is truly important. And how do you thank a sister who visits you when you need her, who stays away when you do not, who "sold" you on MIT, who introduced you to Ashdown House, who helped format your 200+ page document so that you would not have to struggle with Microsoft, and who looks out for you in big and small ways?

Finally, thanks and praise to the Lord Jesus Christ, without whom, ultimately, neither this document, nor this community, would ever have come to pass.

> Unless the Lord builds the house, its builders labor in vain.
> Psalm 127:1

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

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.

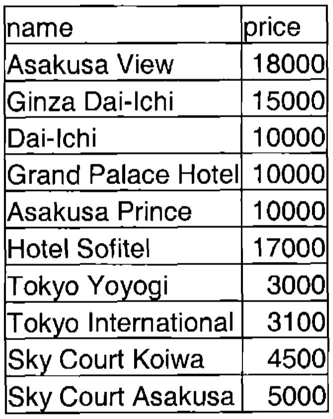
Today, the World Wide Web is a popular way to access the Internet. One group of tools to help people navigate the labyrinth of on-line content are integration services that allow a user to pose rich queries across multiple sites and aggregation services which effectively roll several different sources 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

This thesis 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). This thesis identifies a set of challenges to integration that stem from the problem of attribution (i.e. knowing where data comes from). The challenges embrace a range of technology and policy questions. Therefore, the thesis is divided into two parts. We begin with a technology-based approach to documenting data sources. A formal model of attribution is introduced to support the capability of integrating data from heterogeneous sources. We then expand the scope of our examination from technologies that support data integration to the general issue of data integration regardless of the means for doing so. Policy measures to both limit and support integration based upon where information comes from are considered.

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. Both the formal treatment and the broader policy view draw upon this definition of attribution. Up to this point, we have used the term 'attribution' colloquially, relying upon context to provide the user with an intuition for the term. In Chapter 2, we provide an operational definition for the task of attribution as a list of desiderata for any attribution technology or policy.

Because of society's ever deepening dependence upon streams of data, we have not been the only individuals interested in the integration-attribution problem space. It becomes clear that over time, no small amount of theoretical and empirical research, often in different guises, has already been leveled at the general problem of attribution. Chapter 2 provides a very brief overview of a number of the diverse, perhaps seemingly unrelated research streams. Research approaches and results more similar to our own or upon which we draw heavily are revisited and discussed in greater detail throughout the thesis.

Part 1 proposes one technological approach to addressing attribution-related challenges. We develop a formal model of attribution in the context of the relational data model. Although motivation for this work largely stems from efforts to introduce transparency to the heterogeneous, semistructured environment that is the World Wide Web, we build our theory in the relational context because the relational data model provides firm theoretical grounding and is the foundation for the most widely used commercial database products today.

Part 1 opens with Chapter 3, a high-level tour of the model. Through examples and illustrations, we attempt to provide an intuition for the different concepts and principles that the model aims to characterize. In Chapter 4, we extend our intuitions to a formal model. Our goal in providing a formal model is to offer a consistent framework for interpreting different facets of attribution and understanding how those different dimensions relate to one another. Our formalization is based upon the proof semantics of the domain relational calculus (DRC). A brief review of the specific syntax and semantics assumed is provided.

After presenting the model and some of its properties, we extend the relational algebra in Chapter 5 to support one instance of the model. We consider some general properties of algebraic extensions such as closure and expressiveness. Then we evaluate the degree to which the extended algebra implements the model. Finally, revisiting the example from Chapter 6 that originally motivated our exploration of attribution, we begin a discussion of extensions to our model of attribution.

Part 2 of this thesis returns to the general question of promoting integration while preserving the incentives for producing the underlying data sources. Our technology discussion required a narrow focus on the task of attribution itself. Now, we revisit the broader attribution problem space first introduced in Chapter 2. We consider both traditional and novel measures that judges and legislators have invoked to craft the current policy framework surrounding data integration technologies.

Chapter 7 is a policy analysis. We survey the current policy landscape by revisiting the challenges posed in Chapter 2 from the broader, policy perspective. Then, we review the status quo legal framework addressing those issues, identify the stakeholders, and catalog their respective interests. Chapter 8 is a policy formulation exercise. We begin by clarifying the policy objectives and then redefining 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 9 concludes the thesis with an evaluation that compares our theory of attribution to the desiderata in Chapter Two and that compares our policy formulation to the stakeholder interests in Chapter Seven. As a part of the evaluation, we discuss both limitations of and proposed extensions to this research.

1.2 Scope

This thesis 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. 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.

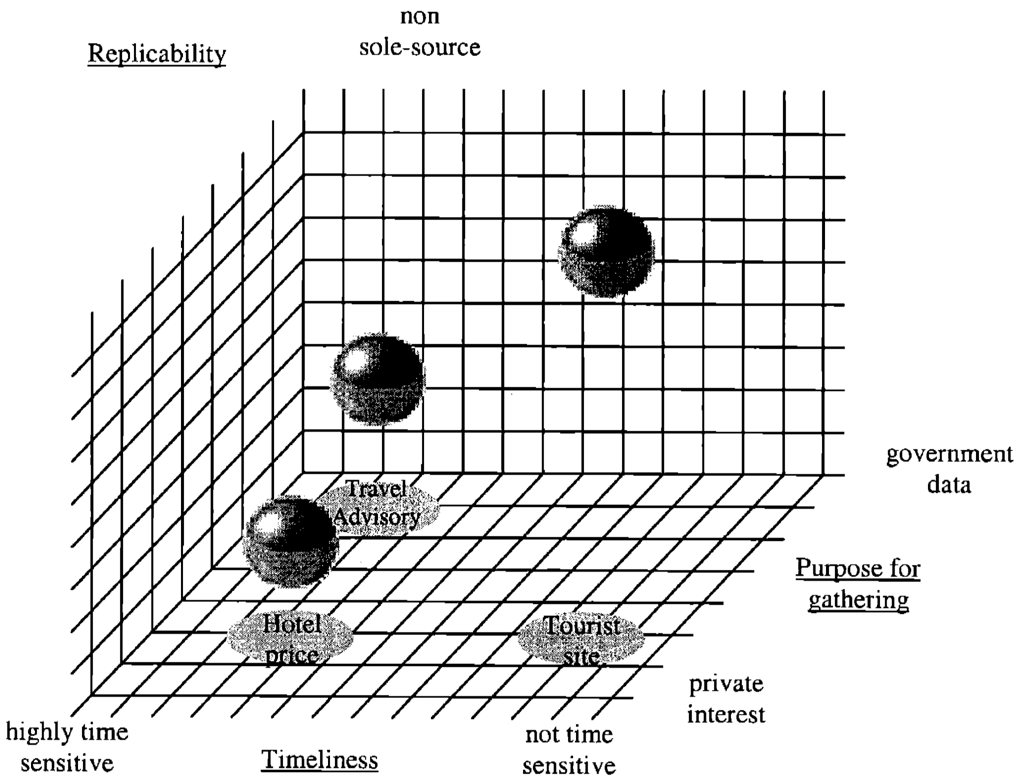


Figure 1.1 Three of the four dimensions of data

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).

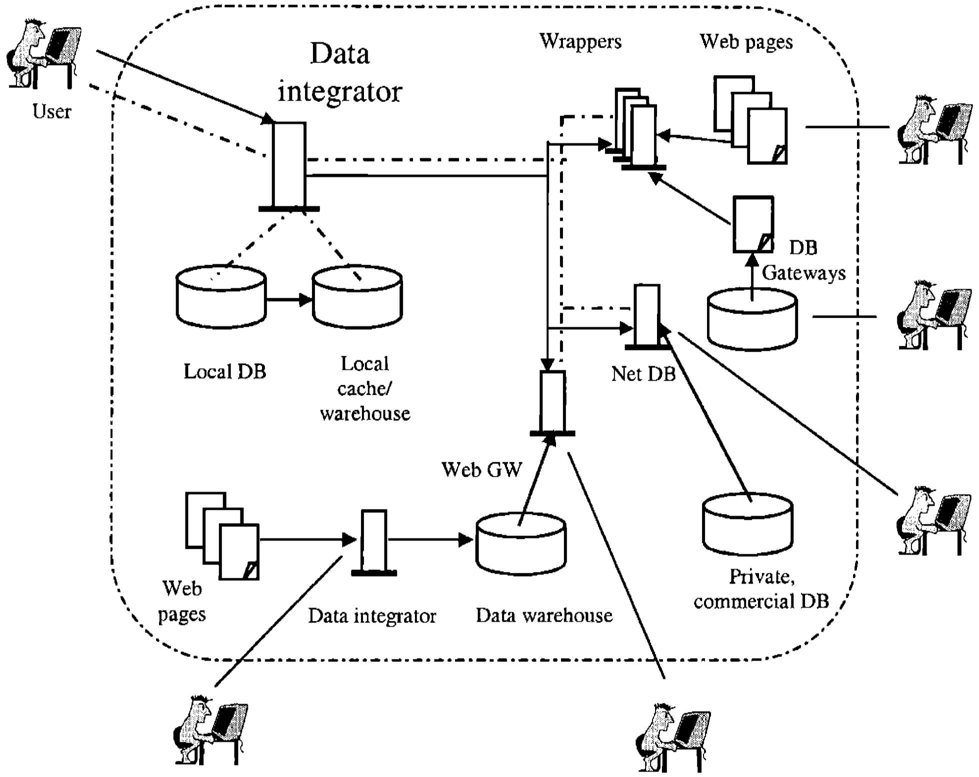


Figure 1.2 Integration architecture

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.

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

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.

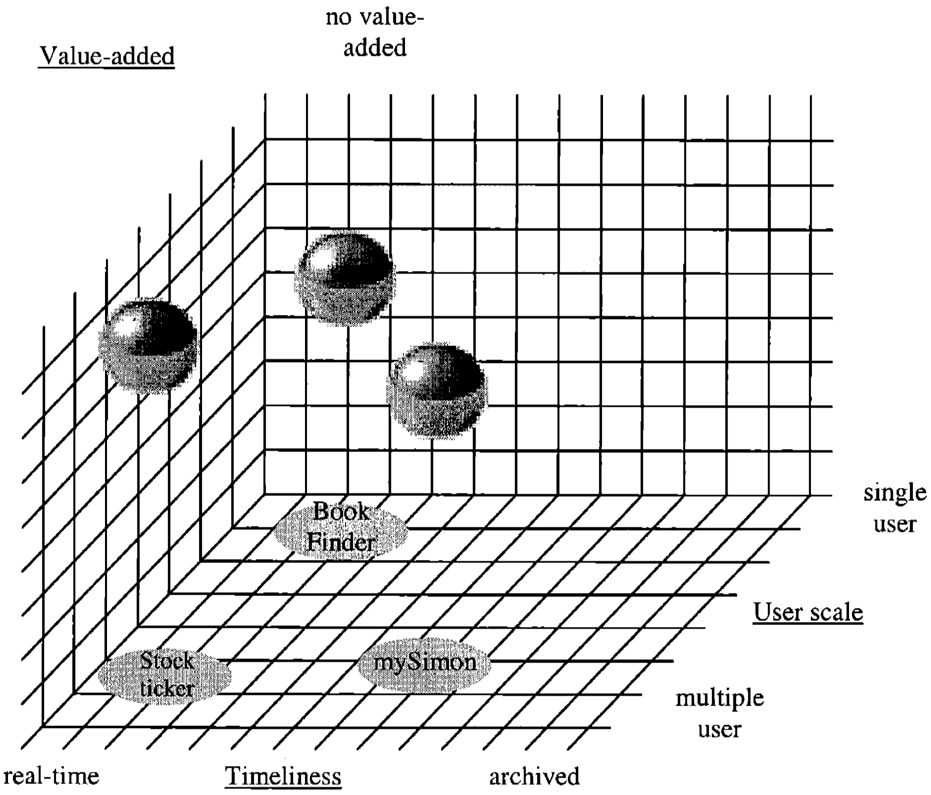


Figure 1.3 Integration strategies

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.

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

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?

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).

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.

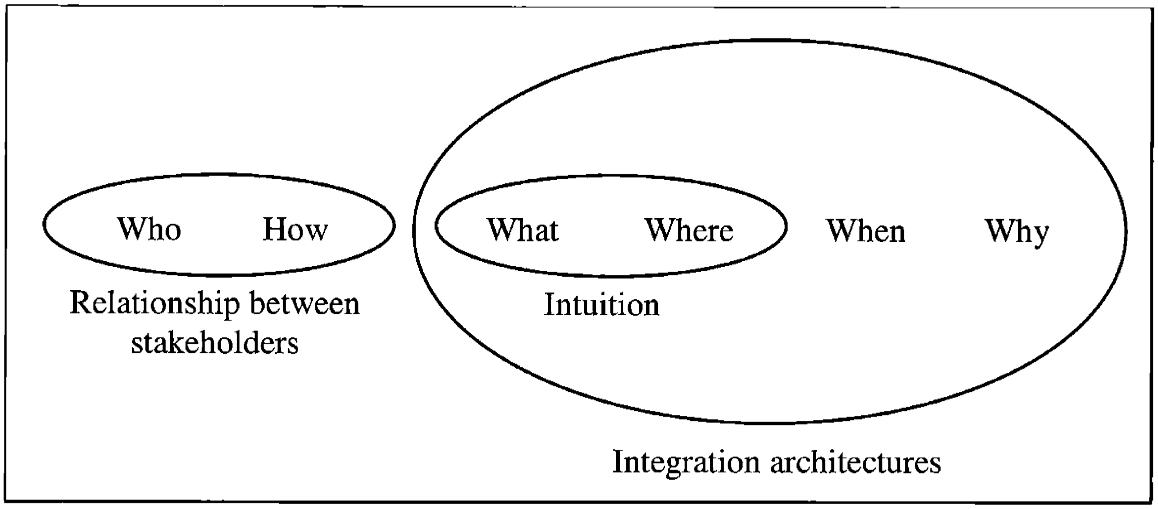


Figure 1.4 The attribution problem space

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 Chapter 1.

Rather than attempting to survey the entire body of related work, we focus on research most similar to our own. Where useful to do so, we attempt to direct the reader to specific application domains or other lines of work that may prove fruitful either for future extensions or to complement that which is presented in this thesis.

Following the structure of the thesis, we first survey technologies to address the attribution problem space and follow that review with policy efforts to treat the same broad topics. For each thread, we discuss related theoretical and pragmatic work.

2.1 Technology approaches to the attribution problem space

We defined the breadth of the problem space in Chapter 1 based upon the dimensions of *who* is gathering and integrating data, *what* data is gathered, *where* the data comes from, *when* the data is collected, *why* or on whose behalf the content is collected, and *how* the integrated collection is used. While there are many technology-based approaches to specific dimensions of the problem (e.g. cookies and Web logs are two approaches to identifying *who*), attribution focuses on drawing the connection between *what* and *where.*

2.1.1 Formal approaches

Research on the relationship between *what* and *where* falls is separable into formal approaches and pragmatic experience. Pragmatic experience is discussed below. Formal approaches in the literature define attribution in one of two ways: the relational algebra and the relational calculus.

The attribution model developed in this thesis was inspired by the Polygen data model, which was first presented in (Wang and Madnick 1990). Though they do not offer a formal definition, Wang and Madnick implicitly define attribution algebraically, as part of a system to assess data quality in heterogeneous data integration. In a Polygen relation, every value has two sets of metadata associated with it. For each result value, input relations are classified into one of three categories: a *source*, an *intermediate*, or irrelevant. The *source* set and *intermediate* set each constitute a heuristic for assessing the quality of a value and the quality of the overall query result. The *sources* for a value in the result are inductively defined as the algebraic input relations that contain those tuples from which said value derives. *Intermediates* are those relations used to evaluate algebraic selection conditions for the query result. Granularity is introduced implicitly. Specific values in the result (fine-grained result granules) are linked to base relations (coarse-grained source granules).

Sadri's work on Information Source Vectors (ISVs) also provides an implicit, algebraic definition of attribution by defining the quality of a tuple in the query result (Sadri 1991; 1994; 1995). Like the Polygen data model, ISVs also classify input relations into one of three roles. ISVs, however distinguish between corroborating and contradictory sources. A source vector, with one slot for every input relation in the database, is associated with every tuple of every base and intermediate relation. The ISV for a result tuple is inductively derived from the ISVs of the algebraic query inputs. Each source vector implicitly corresponds to our notion of *comprehensive* attribution. Because sources are not distinguished from intermediates, Sadri can associate a source vector with every tuple in a relation rather than every value in a relation.

It is worth noting that there exists a host of other works, some of which we will mention in the context of pragmatic approaches to attribution below, that also rely upon implicit, algebraic definitions of attribution. Domain and application specific research in the area of Census data tracking, Geographic Information Systems, and security authorization (Ferber 1991; 1992; Lanter 1991; Lanter and Surbey 1994; Motro 1996; Motro and Rakov 1998; Rosenthal and Sciore 1999a; b; Woodruff and Stonebraker 1997) all determine some meta-characteristic of a value or a tuple in a result based upon the processing of input relations. Some (Woodruff and Stonebraker 1997) define fine-grained lineage, associating result values with input values rather than input relations. Note that we may frame some of the research in probabilistic or temporal databases similarly (Dey, Barron, and Storey 1996; Dey and Sarkar 1996). The probabilities or temporal ranges are a function of the constituent inputs. From the perspective of defining attribution based upon the query processing operations, however, they are all essentially similar.

The research in this thesis builds from earlier work that combines the concept of attribution with a specific metric that derives from the input relations such as data quality or access permissions. We extend the existing literature in several respects. First, we provide an explicit definition of attribution. This definition is couched in terms of the relational calculus and the logical foundations for relational database theory rather than implicitly in the algebra. Second, we refine the concepts of *source* and *intermediate* to distinguish between three types of attribution, *comprehensive*, *source*, and *relevant*, to correspond to different user needs. Third, based upon the formal model we can express equivalence properties for attribution. Finally, we attempt to articulate granularities explicitly and then suggest how the relationship between source and result granules may support subsequent algebraic extensions to reduce the burden of propagating attribution metadata.

In contrast to the implicit algebraic definitions of some of the early work in source tracking, Cui et al. (2000; 2001; 1997 (revised 1999)) provides a formal definition of *lineage*, in terms of the relational algebra. Reflecting their primary application domain, data warehousing, Cui et al. further extend their definition of lineage first to encompass bag semantics and aggregation functions and later to more general classes of transformations (e.g. arithmetic functions in a select clause, grouping tuples, etc.). For the base relational operators, the lineage of a result is recursively defined by the successive application of operators in the query tree. Equivalence properties of lineage are defined. As with Sadri (1991), corresponding to their focus on comprehensive attribution, Cui et al. (1997 (revised 1999)) define attribution for result tuples. Unlike earlier work, however, they focus on "fine-grained" lineage and associate result tuples with input source tuples rather than input relations.

Given our characterization of the attribution problem space, we define three different types of attribution rather than one. Each type of attribution has somewhat different properties with respect to both equivalence and granularity. *Lineage*, as defined in (Cui, Widom, and Wiener 1997 (revised 1999)), corresponds to our concept of *comprehensive* attribution. We also attempt to define the relationship between source and result granules explicitly.

The relational calculus and the relational algebra are equal in their expressiveness. Consequently, neither model is necessarily better than the other for defining attribution. However, as is echoed in the work by Buneman et al. (1998; 2001), the different semantics of calculus queries provides a more direct parallel to languages for querying semistructured data on the Web; and it is the reuse and redistribution exacerbated by the Web that underlies our interest in attribution.

The second category of theoretical approaches builds or borrows from the first-order predicate logic with which the relational calculus is defined. In the relational calculus, queries take the form of expressions on predicates that represent relations. Intuitively, values in a query result are attributable to values from the relational predicates that make the query expression true.

Panorama (Motro 1996) is a system for assessing the quality of data in a query result. Panorama explicitly notes that the same quality assessment(s) might not apply uniformly to all values in the relation (*granularity*). The reliability or completeness of answers are at least partially determined by their contributing sources. Quality properties are thus associated with the subset of tuples in a relation for which the property holds. A tuple subset is proscribed by a *meta-tuple* or select-project view expressed in the relational calculus. A particular property is inherited by a query result if tuples from the corresponding *meta-tuple* provide a true interpretation of the query expression.

Using query expressions to define *meta-tuples* matches our use of expressions to define *source granules*. We extend the intuition one step further to associate source granules with result values rather than tuples. This finer granularity supports three different types of attribution. By contrast, Panorama propagates values based upon our notion of *source attribution* or the specific *meta-tuple(s)* or relations from which result tuples are drawn. Finally, we do not associate source granules with particular properties of the sources, thereby separating the attribution from a specific motivation (e.g. quality, intellectual property, search), leaving the user or application domain to associate their own meta-characteristics.

Buneman et al. (1998; 2001) borrow from the logical intuitions underlying the relational calculus, but generalize the data model to a deterministic semistructured data model. They define both *why* and *where* data provenance for queries (path expressions) in this context. In a separate work, Buneman et al. (2001; 2001) represents the concept of source granules as deep linking into source documents. They also explore the use of key values (in the relational sense) to represent linking into source documents.

The research by Buneman et al. is in many ways most similar to the spirit, approach, and ultimate direction that we aim to pursue in this thesis. Indeed although we structure our formal model in the relational framework to leverage existing results, our initial motivation and long-term aim all along has been to extend the model to semistructured data on the Web. Many of our early intuitions about attribution, such as attribution composition or source and result granularity, stem from this semistructured orientation (Lee, Bressan, and Madnick 1997; 1998).

The semistructured data model is more general than the relational model from which we build in this thesis. However, using the terminology loosely, the *why* provenance for a query on semistructured data is the set of sub-trees that matches the path expression in the same way that we define *comprehensive* attribution as the set of substitutions that provides a true interpretation of a calculus query expression. Indeed (Buneman, Khanna, and Tan 2001) draws upon the same conjunctive query literature that we leverage in exploring equivalence properties (Klug 1988; Sagiv and Yannakakis 1980; Ullman 1989). Similarly, *where* provenance corresponds to our notion of *source* attribution, which in turn stems from the *source* set for every value in a Polygen relation.

Framing our work in the relational calculus, as noted earlier, allows us to borrow directly from the existing literature on equivalence and containment. We are, however, limited to intuitions and observations about the parallels to querying in semistructured environments. We introduce three types of attribution, which better support not only the motivations of the attribution problem space but relate to the relationship between source and result granules. We also treat explicit equality in theta comparisons independently of the natural join. This reflects a distinction in *source* attribution (*where* provenance) relevant to such purposes as intellectual property or remuneration. The natural join suggests that both relations are sources for the join attribute whereas explicit equality indicates that each argument to the equality has only one, distinct source. Finally, we also present an extension to the relational algebra as a mechanism for explicitly propagating attribution metadata in annotations.

2.1.2 Pragmatic approaches

Turning from different formal methods for defining attribution, we next consider pragmatic approaches to providing attribution support in querying and integration.
We can separate pragmatic strategies for managing attribution into eager and lazy approaches. Eager approaches continuously update and propagate attribution metadata as a part of query processing. A'priori evaluation, however, amortizes the cost of attribution maintenance over multiple values in the data set and minimizes response time to requests for attribution. We may also think of eager approaches as bottom-up approaches that recursively maintain attribution values.

By contrast, lazy approaches, which may also be thought of as top-down approaches, begin with a query result and drill backwards to trace sources for specific values only in response to specific requests. Minimal expense is incurred in query processing, but the cost of responding to any single attribution request is much higher. Hybrid models may evaluate the attribution for certain intermediate inputs (e.g. frequently used views) to speed-up response to ex-post, lazy attribution requests.

Early work on extensions to the relational data model were, in part, both motivated by and demonstrated using eager attribution principles. Schek and Pistor (1982) articulated their approach to the non-first normal form in the context of merging information retrieval and database approaches to managing search. In their NF2 model, data values are extended with a relation identifying their source(s) as a means for directing subsequent information retrieval queries for additional data. Their early work echoes an attribution driver identified in Chapter 1, searching for related information.

The Polygen data model (Wang and Madnick 1990), upon which this thesis is based, is another prototypical example of an eager approach to attribution. Wang and Madnick extend the relational data model with two annotations - one each for references to *sources* and references to *intermediates*. Every domain value is therefore a triple and a relation is a finite subset of the Cartesian product of such triples. Polygen extensions to the algebra then update values in the *source* and *intermediate* annotations with each successive application of the corresponding operator. References are relation names. The Polygen model therefore provides attribution for individual result values using relation-level source granules.

A number of projects that calculate and propagate meta-attributes of data (e.g. time stamps, probability, quality, authorization) work in a similar manner. In (Dey, Barron, and Storey 1996; Dey and Sarkar 1996), a tuple is tagged with a probability measure or time stamp, respectively. The preservation of certain algebraic equivalencies is demonstrated and, in the case of the temporal relational algebra, aggregation functions are also considered. Both closure and consistency with the traditional relational algebra are verified. Tuples are tagged similarly with quality specifications in (Motro and Rakov 1998). Algebraic extensions manage metadata propagation from constituent inputs to results. In (Rosenthal and Sciore 1999b), security policies are specified as the manner by which security authorizations are aggregated. For example, the permissions on a specific tuple might be the least upper bound of the permissions on all inputs.

That different projects may calculate meta-characteristics at different levels of granularity is perhaps more a function of the application domain than a limitation of the eager approach. Certain applications (e.g. intelletctual property), may wish to identify the Source of a specific value in a tuple while other uses of attribution may require only tuple-level granularity. The principle distinction between these domain specific approaches and the work in this thesis (as well as the Polygen data model from which this work derives) is the propagation of source meta-characteristics (e.g. quality) rather than source references.

Sadri's (1991; 1994; 1995) work on Information Source Vectors (ISVs) suggests the complementary nature of the two approaches to annotation. The relational data model is extended with an ISV annotation for every tuple. Algebraic extensions update and propagate ISVs for result tuples. The quality of a given tuple is then determined as a function of the corroborating and contradictory sources in the corresponding ISV rather than returning a continuously updated metacharacteristic. Where ISVs are associated with result tuples, the attribution in this work is associated with individual values, thereby supporting distinctions between types of attribution.

In addition to eager approaches that extend the data representation with annotations are eager systems that construct parallel data structures for managing attribution metadata. Panorama is one such system (Motro 1996). In Panorama, annotations on the quality (e.g. soundness, completeness) of tuples in a relation are associated with a *meta-tuple* for the relation. A *meta-tuple* is simply a select-project view defining the subset of tuples to which the metric applies. The set of all metrics applicable to a relation is called a *meta-relation*. Queries on relations are paralleled by operations on the corresponding *meta-relation*.

Where eager approaches propagate data continuously, lazy approaches minimize the ex-ante cost of maintaining attribution. A minimum amount of information is stored. Only when a specific request is initiated, is the attribution for a result calculated.

In his work to support data integration and reuse in Geographic Information Systems (GISs), Lanter maintains GIS metacharacteristics in a parallel data structure (Lanter 1991; Lanter and Surbey 1994). Where algebraic operators in the relational model process relational tuples, GISs process *layers*. Lanter defines a frame-based representation to capture layer-level metacharacteristics including data transformations. Operations on layers are paralleled by the updates to the corresponding knowledge-base tracking GIS processing. Specific metacharacteristics are therefore associated with each layer in the manner of tuple-level result granules. The lineage for a result is generated by tracing backwards through the frames associated with each successive processing step.

Like Lanter's system for Geographic Information Systems, Woodruff and Stonebraker (1997) define a system to trace data lineage. Unlike Lanter's layer-granularity that documents metacharacteristics at the level of a data set, Woodruff and Stonebraker register data transformations and their inverses. The inverses allow users to regenerate specific base level data inputs to the transformation process. Original data values are calculated iteratively by unfolding successive operations. The result is fine-grained lineage that traces from a value in the result to the source input values rather than merely linking result sets to their constituent inputs.

Cui et al. (2001) investigates lineage for general data transformations in the spirit of (Woodruff and Stonebraker 1997). However, it is their earlier work tracing relational queries, described in (Cui and Widom 2000; Cui, Widom, and Wiener 1997 (revised 1999)), that our extended algebra is most similar to. Assuming a canonical form of an algebraic query tree, Cui and Widom algorithmically construct a tracing query that, for a given result tuple, returns the input tuples. The algorithm works by essentially projecting the result tuple as query constraints down the algebraic query. The resulting *lineage* is transitive over intermediate results and through querying on views.

Although the technique does not strictly require maintaining meta information, as used in eager approaches, it is possible to achieve greater efficiency in lazy attribution processing by utilizing eager approaches in a limited manner. Cui et al. (1997 (revised 1999)) discover significant improvement in lazy performance by storing auxiliary views, which we might equate with eager evaluation of attribution metadata for intermediate query results. Maintaining a minimal amount of metadata with query processing also enables Cui et al. to trace backwards through aggregation functions.

We adopt an annotation approach to managing attribution metadata. Based upon our formal definition of attribution and our articulation of granularity, we redefine the extended relational operators to support the formal definition of attribution. Unlike some of the approaches that extend the relational model, we show how general properties of the algebra, such as closure, are preserved. Moreover, unlike approaches that rely upon implicit definitions, we show how the algebraic extensions indeed support our logical intuitions about the different interpretations of attribution. Although the algebra tracks source granules at the granularity of relation names, it is a straightforward extension to consider variable granularity using expressions as in Panorama (Motro 1996) rather than relations (Sadri 1991) or explicit source tuples (Cui, Widom, and Wiener 1997 (revised 1999)).

Annotations in a bottom-up manner seems the most general approach for addressing the myriad interests that we initially identified in attribution. Certainly systems designed with specific goals in mind might prefer one particular approach over another. Moreover, the top-down query tracing implemented by Cui et al. is similar in spirit to how Panorama associates result granules with source granules and how we project substitutions onto intermediate relational predicates in attribution composition.

Where *meta-tuples* in Panorama or the metadata in other systems to document data probabilities, quality, or authorization (Dey, Barron, and Storey 1996; Dey and Sarkar 1996; Motro 1996; Motro and Rakov 1998; Rosenthal and Sciore 1999a) are explicitly associated with specific metrics, we define attribution only as the association between source and result granules. Doing so allows us to define different types of attribution and to parametrize attribution with different functions for quality, intellectual property, or search metrics as the need arises.

2.2 Related policy approaches

While the formal and pragmatic technologies reviewed above address 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 expanded the scope of this research to explore 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 on particular on policy approaches similar to our own.

Much of the research literature on 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 Chapter 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 Chapter 7. In Chapter 7, we provide a comprehensive review of the status quo policy approach to questions of *who*, *what*, *where*, *when*, *why* and *how*.

2.2.1 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.

Related work

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

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

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.

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.

2.2.2 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.

2.2.2.1 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.

[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.

RELATED WORK

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 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.

2.2.2.2 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.

We adopt a similar methodology to that found in much of the literature on the economics of information technologies and intellectual property. Building from a transactions cost approach, we apply both a game theoretic view and an Entitlements perspective (Calabresi and Melamed 1972; Gordon 1992). However, in focusing exclusively on database integration, we offer some new perspectives. First, as noted above, we argue in Chapter 8 that databases are a distinctive form of intellectual property. Information technologies affect the transactions costs associated with database (re)use and (re)distribution in ways different from high authorship works. Second, the market for data is not homogeneous. Chapter 7 articulates several different market models for data (re)use. As a consequence, Perritt's cost equations lead to novel conclusions. As noted earlier, it is a consideration of both technologies and economics underlying the industry, as well as technology and policy alternatives for protection, that makes our policy analysis and policy formulation unique

3 Attribution intuitions

In Chapter 1, we provided some rough boundaries about the attribution problem space and some desiderata for a formal approach to that space. Here, we begin Part 1 of the thesis. Beginning with Chapter 3 and extending through Chapter 6, we develop a model for attribution. Although we make the model formal in Chapter 4, we begin in this Chapter by attempting to provide the intuitions behind the features and properties of our proposed model. The intuitions are intended to connect the reader from the problem space defined in Chapter 1 to the formalisms in Chapter 4. After presenting the model, we operationalize one instance of the model as an extension to the relational algebra. Finally, we consider how the model might apply in the emerging semi-structured data environment.

Throughout this Chapter and the remainder of this thesis, we couch many of our examples in the context of the relations listed in Table 3.1. The six relations in Table 3.1 represent a number of separate (Web accessible) data sources concerning lodging and tourist attractions in Tokyo, Japan. The relation hotels(HNAME, ROOM, PRICE) lists hotels in Tokyo along with a minimum price for rooms in the ROOM category. The relation sites(SNAME, REGION) identifies tourist attractions in Tokyo along with the general vicinity where the attraction is located. The three relations roughguides(HNAME, PRICE, STATION, PHONE); jyh(HNAME, PRICE, STATION, PHONE, FAX); and hostels(HNAME, PRICE, STATION) all provide listings of youth hostels or other low-budget lodging in Tokyo. The attribute STATION identifies the closest rail station to the associated lodging. regions(HNAME, REGION) provides the general geographic location of selected Tokyo hotels. Though the model is developed in the DRC, for readability, the examples in this chapter are posed in English, SQL, and the calculus.

3.1 The meaning of attribution

This theory of attribution is based upon the domain relational calculus (DRC), a logical formalism for representing and evaluating relations between data domains. We build our model in this environment because, while our motivation is heavily influenced by the rapid evolution of data integration on the World Wide Web, most of what is known today about managing and manipulating data is rooted in relational terms. The calculus is also the foundation for SQL, one of the most widely recognized and used standards for querying and managing information. In theoretical terms, then, the calculus will allow us to be precise about our observations and intuitions. Pragmatically, much of the data being used today, even that accessible over the Web, is still managed and manipulated using relational tools built on the calculus.

hotels

HNAME	ROOM	PRICE
Asakusa View	single	18000
Asakusa View	double	20000
Ginza Dai-Ichi	single	15000
Ginza Dai-Ichi	double	25000
Imperial Hotel	single	34000
Imperial Hotel	double	39000
Dai-Ichi	single	10000
Dai-Ichi	double	80000
Grand Palace Hotel	single	10000
Grand Palace Hotel	double	31000
Asakusa Prince	single	10000
Asakusa Prince	double	42000
Hotel Sofitel	single	17000
Hotel Sofitel	double	22000

sites

SNAME	REGION
Imperial Palace	Hibiya
Tourist Information Center	Hibiya
Tsukiji Fish Market	Hibiya
Hama Rikyu Garden	Tsukiji
Sensoji Temple	Tsukiji
Nakamise Dori	Asakusa
Ameya Yokocho	Asakusa
Ueno Park	Ueno
Tokyo National Museum	Ueno
Yanaka	Ueno
Meji Jingu Shrine	Ueno

roughguides

HNAME	PRICE	STATION	PHONE
Sky Court Asakusa	5000	Asakusa	81-3-3672-4411
Hotel Pine Hill	10000	Ueno-Hirokoji	81-3-3822-2251
Sawanoya Ryoken	5000	Nezu	81-3-3847-4477
Hotel Top Asakusa	7000	Asakusa	81-3-3822-1611
Ryokan Shigetsu	7000	Asakusa	81-3-3843-2345

jyh

HNAME	PRICE	STATION	PHONE	FAX
Tokyo Yoyogi	3000	Sangubashi	81-3-3467-0163	81-3-3467-9417
Tokyo International	3100	Iidabashi	81-3-3235-1107	81-3-3267-4000
Sky Court Koiwa	4500	Koiwa	81-3-3672-4411	81-3-3672-4400
Sky Court Asakusa	5000	Asakusa	81-3-3672-4411	81-3-3875-4941

hostels

HNAME	PRICE	STATION
Tokyo Yoyogi	3000	Sangubashi
Tokyo International	3100	Iidabashi
Sky Court Koiwa	4500	Koiwa
Sky Court Asakusa	5000	Asakusa
Hotel Pine Hill	10000	Ueno-Hirokoji
Sawanoya Ryoken	5000	Nezu
Hotel Top Asakusa	7000	Asakusa
Ryokan Shigetsu	7000	Asakusa

regions

HNAME	REGION
Hotel Sofitel	Ueno
Katsutaro	Ueno
Dai-Ichi Hotel	Hibiya
Imperial Hotel	Hibiya
Asakusa View	Asakusa

Table 3.1 Data for examples

The interpretation of a calculus expression is the set of variable substitutions that correspond to facts in the database and make the formula of the expression true (Maier 1983). In the most general sense, we express attribution in terms of the substitutions that make the interpretation of the expression true.

Example 3.1 Intuition for attribution
Q1. Based upon the database of Table 3.1, we might ask: What are the names of all known lodging establishments in Tokyo, Japan? We could answer this question by considering the union of a query on the relation hotels and a query on the relation hostels.

```
SQL 1.1   select HNAME from hotels
          union
          select HNAME from hostels
```

DRC 1.1 {HNAME | hotels(HNAME, ROOMS, PRICE) v hostels(HNAME, PRICE, STATION)}

The query result is:

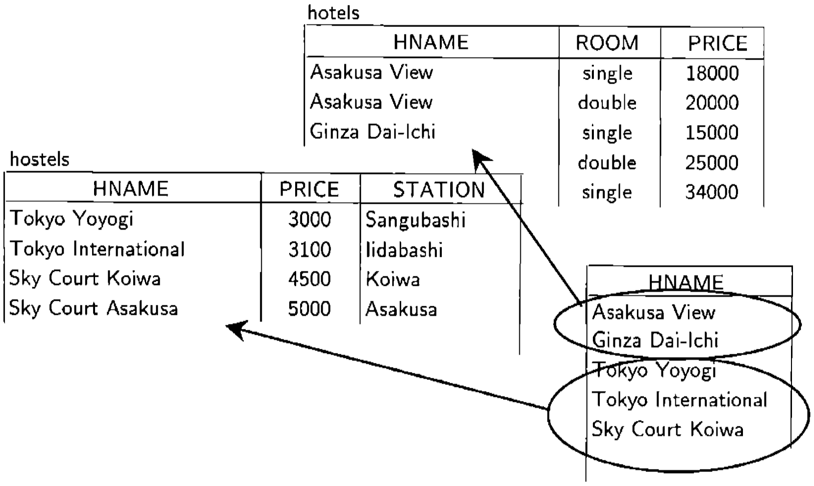
HNAME
Tokyo Yoyogi
Tokyo International
Sky Court Koiwa
Sky Court Asakusa
Hotel Pine Hill
Sawanoya Ryoken
Hotel Top Asakusa
Ryokan Shigetsu
Asakusa View
Ginza Dai-Ichi
Imperial Hotel
Dai-Ichi
Grand Palace Hotel
Asakusa Prince
Hotel Sofitel

Table 3.2 Lodging establishments in Tokyo, Japan

Some of the substitutions that provide true interpretations include the following:
<*f*("Asakusa View"/HNAME, "single"/ROOMS, 18000/PRICE)>;
<*g*("Tokyo Yoyogi"/HNAME, 3000/PRICE, "Sangubashi"/STATION)>;
<*g*("Sky Court Asakusa"/HNAME, 5000/PRICE, "Asakusa"/STATION)>;
<*f*("Dai-Ichi"/HNAME, "double"/ROOMS, 10000/PRICE)> □

If we further represent relations as *sources* for data, we can talk about different roles that sources play based upon the substitutions (facts) from each source used to interpret the expression. Future references to '*sources*' in this chapter will refer to the relations containing the facts which, when substituted into the query expression, produce a true interpretation.

Example 3.2 Intuition for a "source"
Given the substitutions for Q1 in Example 3.1, the corresponding *sources* are: relation hotels and relation hostels. We depict this intuition in Figure 3.1. From the answer, a list of HNAME, we can trace backwards to the corresponding input relations. □



hotels

HNAME	ROOM	PRICE
Asakusa View	single	18000
Asakusa View	double	20000
Ginza Dai-Ichi	single	15000
	double	25000
	single	34000

hostels

HNAME	PRICE	STATION
Tokyo Yoyogi	3000	Sangubashi
Tokyo International	3100	Iidabashi
Sky Court Koiwa	4500	Koiwa
Sky Court Asakusa	5000	Asakusa

HNAME
Asakusa View
Ginza Dai-Ichi
Tokyo Yoyogi
Tokyo International
Sky Court Koiwa

Figure 3.1 Intuition for a "source"

We saw in Chapter 1 that there may be different motivations for or interests in attribution. Accordingly, our theory defines three explicit types of attribution: comprehensive, source only, and relevant. *Comprehensive* attribution identifies everything that was used to evaluate an expression. It identifies every source that was consulted. Certainly from the perspective of remuneration, comprehensive attribution is in the interests of data providers. From a data quality perspective, comprehensive attribution provides a measure of completeness regarding the answer to a query.

Source attribution, by contrast, recognizes the difference between "supporting material" and the actual facts. *Source* attribution identifies the specific relations from which a query result is drawn. We use the metaphor of a footnote in a text citation. Unlike the *comprehensive* listing of references in a bibliography, a footnote identifies author, title, and page number for a specific fact, figure, or quotation. Certainly for intellectual property purposes, source attribution is critical. Moreover, as measure of quality distinct from that of *comprehensive* attribution, we may use the credibility of a given source to label the veracity of the data from that source. Finally, knowing the specific source of a data item provides us with a starting point for seeking additional, related information.

Relevant attribution constitutes a subset of *comprehensive* attribution. Given a specific result, the *relevant* attribution identifies the subset of *comprehensive* references that are associated with the *source* attribution of a particular query. For example, the *comprehensive* list of references in this thesis numbers over 250 separate works. However, our treatment of negation in Chapter 4 draws from work by Sagiv and Yannakakis (Sagiv and Yannakakis

1980). However, we found this reference through a series of other works (Abiteboul, Hull, and Vianu 1995; Ullman 1989). *Relevant* attribution therefore traces the supporting material used to arrive at a single query. In SQL terms, we may think of *relevant* sources as those used in evaluating selection conditions.

In simple queries, the *comprehensive, source,* and *relevant* attribution may look identical. As query complexity increases, however, particularly in the light of the data environment of the Web, such distinctions may become increasingly important in parsing the attribution problem space.

Example 3.3 Types of attribution

Q2. Consider the query where we ask for all hotels by the Imperial Palace in Tokyo, Japan. Based upon the hypothetical database of Table 3.1, we have:

```
SQL 2.1   select HNAME
          from hotels, regions, sites
          where sites.SNAME = "Imperial Palace"
          and sites.REGION = regions.REGION
          and hotels.HNAME = regions.HNAME

DRC2.1    {HNAME | regions(HNAME, REGION) ∧ sites("Imperial Palace", REGION) ∧
          hotels(HNAME, ROOMS, PRICE)}
```

The substitutions include (but are not limited to):

```
<f("Imperial"/HNAME, "Hibiya"/REGION, "Imperial Palace"/SNAME, "single"/ROOMS,
      34000/PRICE)>;
<f("Imperial"/HNAME, "Hibiya"/REGION, "Imperial Palace"/SNAME, "double"/ROOMS,
      39000/PRICE) >;
<f("Dai-Ichi"/HNAME, "Hibiya"/REGION, "Imperial Palace"/SNAME, "single"/ROOMS,
      10000/PRICE) >;
<f("Dai-Ichi"/HNAME, "Hibiya"/REGION, "Imperial Palace"/SNAME, "double"/ROOMS,
      80000/PRICE) >;
```

Now, consider the relations from where these substitutions are drawn. The different substitutions are drawn from three different relations. Therefore, the *comprehensive* attribution includes these three relations. But, not all of the relations in the FROM clause of the SQL query are used to provide answers. As illustrated in Figure 3.2, some sources are used to evaluate selection conditions rather than provide selection attributes. In particular, the HNAME attribute that constitutes the query result appears in only two of the queried relations. Thus, the *source* attribution includes only two relation names. Finally, because the relation sites is used in evaluating selection conditions, we include it in the *relevant* attribution.[6]

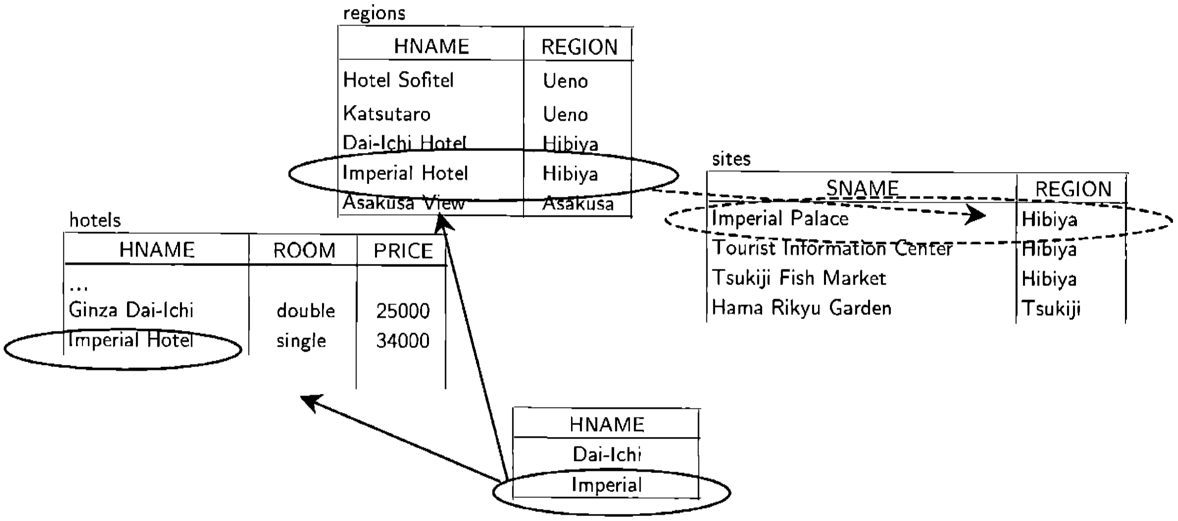
comprehensive attribution

{<regions; sites; hotels>}

[6] For an example where *comprehensive, source* and *relevant* attribution are all different for the same query expression, see Example 3.7 where we consider the Union query operator.

source attribution
{<regions; hotels>}

relevant attribution
{<regions; sites; hotels>} □



regions

HNAME	REGION
Hotel Sofitel	Ueno
Katsutaro	Ueno
Dai-Ichi Hotel	Hibiya
Imperial Hotel	Hibiya
Asakusa View	Asakusa

sites

SNAME	REGION
Imperial Palace	Hibiya
Tourist Information Center	Hibiya
Tsukiji Fish Market	Hibiya
Hama Rikyu Garden	Tsukiji

hotels

HNAME	ROOM	PRICE
...		
Ginza Dai-Ichi	double	25000
Imperial Hotel	single	34000

HNAME
Dai-Ichi
Imperial

Figure 3.2 Example of source attribution

3.2 Properties of attribution

A specific challenge to any theory of attribution is treatment of multiple derivations. Data may derive from many different sources and/or diverse combinations of sources. Accordingly, this theory identifies several distinct categories of multiple derivations and provides an explicit treatment for each. We loosely separate multiple derivations into two categories. Case 1 concerns multiple queries that (appear to) achieve the same result. Think of this as asking the same question in two different ways. For example, "What is for dinner" rather than "What are we eating tonight?" Case 2 addresses a single query that may produce the same answer from more than one source. For example, to discover all the hostels in Tokyo, Japan, you might combine the results from looking in both a Japanese travel guide and an international youth hostel guide. Some entries might be listed in both places.

Case 1, multiple queries that (appear) to achieve the same result, is further separated into three classes: weak equivalence, strict equivalence, and composition. Weak equivalence, in a colloquial sense, refers to queries that, perhaps in some circumstances, appear as if they should be equivalent yet are not logically equivalent and therefore vulnerable to incomplete data or other contextual limitations (Ullman 1989).

Example 3.4 Weak equivalence
Q3. Consider the query that asks for all hotels in Tokyo, Japan. Given only the schemas for the relations in Table 3.1, we might conclude that there are at least three different ways to list hotels in Tokyo.

```
SQL 3.1   select HNAME from regions
SQL 3.2   select HNAME from hotels
SQL 3.3   select HNAME from regions, hotels where hotels.HNAME = regions.HNAME
```

Unfortunately, as is often the case in real tables, our example data relations are incomplete. There are a number of dangling tuples (Ullman 1989). The incompleteness is especially apparent when we consider the results from each of SQL 3.1 – 3 as noted in Table 3.3. □

HNAME
Hotel Sofitel
Katsutaro
Dai-Ichi Hotel
Imperial Hotel
Asakusa View

SQL 3.1

HNAME
Asakusa View
Ginza Dai-Ichi
Imperial Hotel
Dai-Ichi
Grand Palace Hotel
Asakusa Prince
Hotel Sofitel

SQL 3.2

HNAME
Hotel Sofitel
Imperial Hotel
Asakusa View

SQL 3.3

Table 3.3 Weak equivalence

In principle, it seems only reasonable that the data in a database should be somehow complete and internally consistent. Yet, different tables appear to list different hotels even though they all purport to list hotels in Tokyo, Japan. Though a subject studied in the query optimization literature, we do not consider weak equivalents to constitute multiple derivations and so treat them as distinct queries and say nothing more about them.

Strict equivalence refers to the characteristic that two queries produce the same result given the same database.[7] We introduce the modifier "strict" to emphasize the fact that the multiple queries use the same data sources.

Example 3.5 Strict equivalence
Consider again Q2 which we can express in the DRC as

DRC 2.1 {HNAME | regions(HNAME, REGION) ∧ sites("Imperial Palace", REGION) ∧
hotels(HNAME, ROOMS, PRICE)}

DRC 2.2 {HNAME | regions(HNAME, REGION) ∧ sites("Imperial Palace", REGION) ∧
hotels(HNAME, ROOMS, PRICE) ∧ regions(AHOTEL,AREGION)}

[7] We refer to the more formal definition of equivalence based upon containment in Chapter 4 (Ullman 1989).

A substitution for DRC2.1 might look like:

<*f*("Dai-Ichi"/HNAME, "Hibiya"/REGION, "Imperial Palace"/SNAME, "single"/ROOMS,
10000/PRICE)>

A substitution for DRC2.2 might look like:

<*f*("Dai-Ichi"/HNAME, "Hibiya"/REGION, "Imperial Palace"/SNAME, "single"/ROOMS,
10000/PRICE, "Asakusa View"/AHOTEL, "Asakusa"/AREGION)>

Note the similarities between the different substitutions. There are more variables in DRC2.2, yet there is a consistency between the substitutions in DRC2.1 and DRC 2.2. Moreover, our intuitions about attribution are the same for both queries.

comprehensive attribution:
{<regions; sites; hotels>}

source attribution
{<regions; hotels>}

relevant attribution
{<regions; sites; hotels>}

In particular, for the case of strict equivalence, none of the data sources is defined in terms of other available sources. □

Example 3.6 Defining a source in terms of other sources

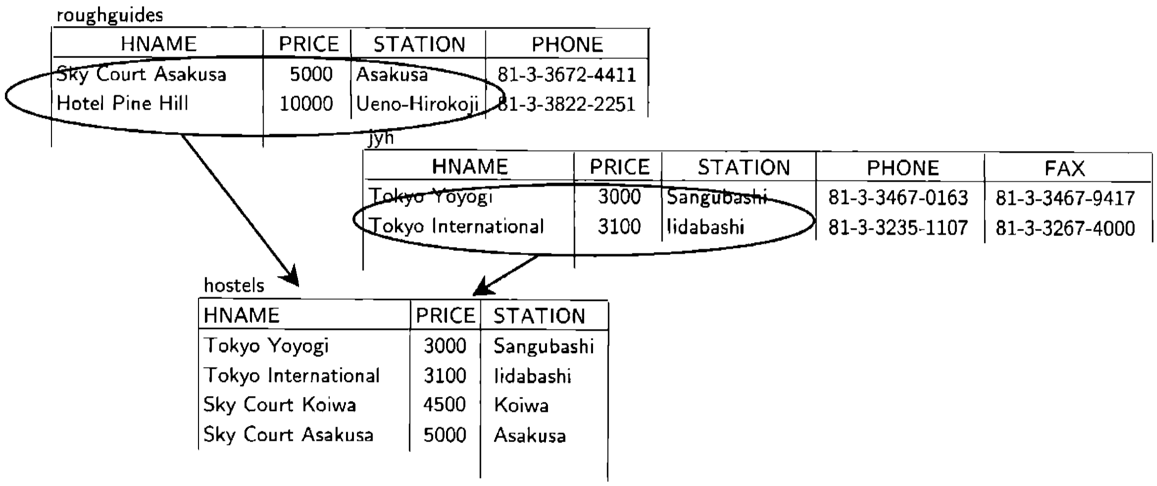
Q4. Consider the query for all hostels in Tokyo, Japan

SQL 4.1 select * from hostels

The reliance of multiple intermediaries upon the same underlying base sources is not always immediately apparent, however. For example, we define relation hostels in terms of information from Japan Youth Hostels Association (relation jyh) and Rough Guide Travel (relation rg). The relationship is depicted in Figure 3.3. Data is taken from the constituent relations to construct a new relation.

```
SQL 4.2   select HNAME, PRICE, STATION from jyh
          union
          select HNAME, PRICE, STATION from rg □
```

In focusing only on strict equivalence, we borrow from the query optimization literature to arrive at the result that the attributions for equivalent select, project, join queries involving theta inequality and natural join are, in some sense, the same. Attribution equivalence is evident in Example 3.5 where, although DRC2.2 has more variables and predicates, there is the sense that there is no extra information gained. We make this intuition explicit when we define attribution equivalence more formally in Chapter 4. However, attribution equivalence is lost for complete and source attribution when we consider queries with union.



roughguides

HNAME	PRICE	STATION	PHONE
Sky Court Asakusa	5000	Asakusa	81-3-3672-4411
Hotel Pine Hill	10000	Ueno-Hirokoji	81-3-3822-2251

jyh

HNAME	PRICE	STATION	PHONE	FAX
Tokyo Yoyogi	3000	Sangubashi	81-3-3467-0163	81-3-3467-9417
Tokyo International	3100	Iidabashi	81-3-3235-1107	81-3-3267-4000

hostels

HNAME	PRICE	STATION
Tokyo Yoyogi	3000	Sangubashi
Tokyo International	3100	Iidabashi
Sky Court Koiwa	4500	Koiwa
Sky Court Asakusa	5000	Asakusa

Figure 3.3 Views: defining sources from other sources

Example 3.7 Attribution equivalence breaks down under union
Consider again Q3, which we defined as all hotels in Tokyo, Japan.
We originally answered this question with

```
SQL 1.1   select HNAME from hotels
          union
          select HNAME from hostels
```

Perversely, we might equally answer the query this way:

```
SQL 1.2   select HNAME from hotels
          union
          select HNAME from hostels
          union
          select HNAME
          from hotels, regions, sites
          where sites.SNAME = "Imperial Palace"
          and sites.REGION = regions.REGION
          and hotels.HNAME = regions.HNAME
```

SQL 1.2 corresponds to:

DRC1.2 {HNAME | hotels(HNAME, ROOMS, PRICE) $\vee$ hostels(HNAME, PRICE, STATION, PHONE) $\vee$ (regions(HNAME, REGION) $\wedge$ sites("Imperial Palace", REGION) $\wedge$ hotels(HNAME, ROOMS, PRICE))}

Compare the attribution between DRC1.1 and DRC1.2 as listed in Table 3.4. In particular, the *source* attribution takes into account the sources used in evaluating each disjunct.

However, the third disjunct is arguably irrelevant because any answer in the third disjunct appears also in one of the first two disjuncts.

Returning briefly to Example 3.3, we see that SQL 1.2 and its associated DRC help highlight the intuition behind the different types of attribution. First, consider comprehensive attribution. Each disjunct represents a distinct alternative for satisfying the DRC1.2. However, taken together, every relational predicate in the query expression plays some role in evaluating the result. The reader will note that the query only asks about hotel names (HNAME), however. Hotel names are only found in four of the five relations used in the expression. If, for example we wanted to verify the spelling of a particular hotel name, there would be no reason to return to relation (sites). That relation does not list any hotel names. Source and comprehensive attribution are therefore distinct. Finally, as observed earlier, the third disjunct in DRC1.2 is contained (or subsumed) by the first two disjuncts. As a consequence, the third disjunct cannot impact the query results and so we omit relations from the third disjunct. The third disjunct is not relevant.[8]

□

	DRC1.1	DRC1.2
comprehensive	{<hotels>; <hostels>}	{<hotels>; <hostels>; <hotels; regions; sites>}
source	{<hotels>; <hostels>}	{<hotels>; <hostels>; <hotels; regions>}
relevant	{<hotels>; <hostels>}	{<hotels>; <hostels>}

Table 3.4 Attribution equivalence with union

The "strict" condition contrasts the third class of queries: "composition," where sources are defined in terms of one another. We saw in Example 3.6 what it means for a source to be defined in terms of other sources, often referred to as views.[9] Composition addresses the situation where a query can either be composed on a view or expressed strictly in terms of the original sources underlying any view definition.

Example 3.8 Query composition

Q5. Consider a query for all lodging (hostels and hotels) around the Nakamise Dori. Based upon Example 3.6, we know that we can express the query in terms of the relations for hostels and hotels:

[8] It is worth emphasizing that while relations may prove irrelevant, they are not without value. As in comprehensive attribution, we may use equivalent derivation paths to increase our confidence in a particular result. Although outside the scope of this work, we may also consider the role of disjuncts which are, in principle, contained but may contain contradictory information (Sadri 1991).

[9] In the relational context, relations defined in terms of other relations are often referred to as views. In the literature on databases and logic, such relations are referred to as intentional databases or IDB (Ullman 1989).

ATTRIBUTION INTUITIONS

```
SQL 5.1   select HNAME
          from hotels, regions, sites
          where hotels.HNAME = regions.HNAME
          and regions.REGION = sites.REGION
          and sites.SNAME = "Nakamise Dori"
          union
          select HNAME
          from hostels, sites
          where hostels.STATION = sites.REGION
          and sites.SNAME = "Nakamise Dori"
```

But, if we know in advance, as we know now, that relation hostels itself gathers information from elsewhere, we can also express the query in terms of the underlying data sources jyh and roughguides (rg) as:

```
SQL 5.2   select HNAME
          from hotels, regions, sites
          where hotels.HNAME = regions.HNAME
          and regions.REGION = sites.REGION
          and sites.SNAME = "Nakamise Dori"
          union
          select HNAME
          from jyh, sites
          where jyh.STATION = sites.REGION
          and sites.SNAME = "Nakamise Dori"
          union
          select HNAME
          from rg, sites
          where rg.STATION = sites.REGION
          and sites.SNAME = "Nakamise Dori"
```

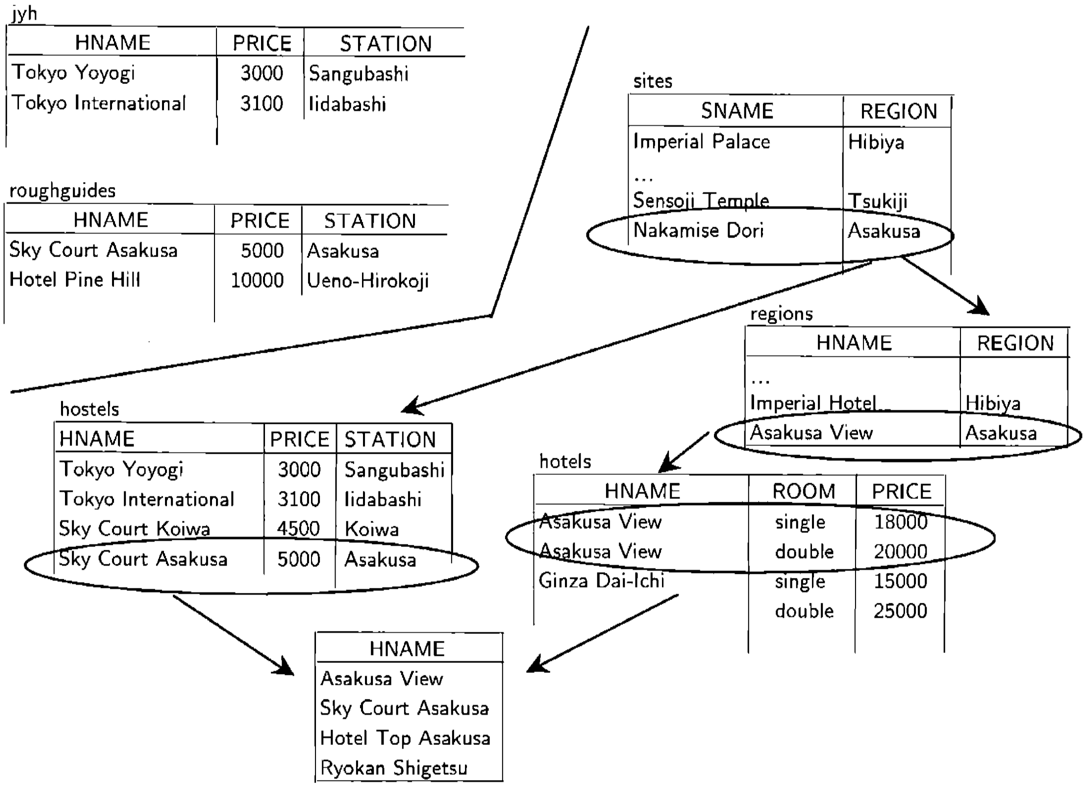
We depict the intuition behind composition in Figure 3.4. SQL 5.1 uses only two relations in the second disjunct (hostels and sites). It is as if relations jyh and roughguides are hidden and inaccessible. Relation hostels then constitutes a view on the underlying sources. The attributions for both queries is shown in Table 3.5. □

	SQL 5.1	SQL 5.2
comprehensive	{<hotels; regions; sites>; <hostels; sites>}	{<hotels; regions; sites>; <jyh; sites>; <rg; sites>}
source	{<hotels; regions>; <hostels>}	{<hotels; regions>; <jyh>; <rg>}
relevant	{<hotels; regions; sites>; <hostels; sites>}	{<hotels; regions; sites>; <jyh; sites>; <rg; sites>}

Table 3.5 Attribution with composed queries

By definition, the query results of equivalent, composed queries are the same. The attributions, however, can be quite different. This seems entirely correct. In the context of distributed, heterogeneous information sources, such as the Web today where data is frequently reused and redistributed, it is not unreasonable to cite an integrator as a source. Factors that are beyond the scope of this thesis, such as reputation or trust may suffice as a proxy for or even improve the perceived quality of the data.[10]

That some needs may be met by attributing to an intermediary source, however, does not preempt the need to know more. We might still wish to look beyond the integrator, unfolding layers of reuse and redistribution back to the underlying initial data sources. We therefore propose an algorithm for unfolding an attribution by recursively attributing values in the intermediary. Based upon this algorithm, we conclude that we can compose an attribution in the same way that we compose relational queries.



jyh

HNAME	PRICE	STATION
Tokyo Yoyogi	3000	Sangubashi
Tokyo International	3100	Iidabashi

roughguides

HNAME	PRICE	STATION
Sky Court Asakusa	5000	Asakusa
Hotel Pine Hill	10000	Ueno-Hirokoji

sites

SNAME	REGION
Imperial Palace	Hibiya
...	
Sensoji Temple	Tsukiji
Nakamise Dori	Asakusa

regions

HNAME	REGION
...	
Imperial Hotel	Hibiya
Asakusa View	Asakusa

hostels

HNAME	PRICE	STATION
Tokyo Yoyogi	3000	Sangubashi
Tokyo International	3100	Iidabashi
Sky Court Koiwa	4500	Koiwa
Sky Court Asakusa	5000	Asakusa

hotels

HNAME	ROOM	PRICE
Asakusa View	single	18000
Asakusa View	double	20000
Ginza Dai-Ichi	single	15000
	double	25000

HNAME
Asakusa View
Sky Court Asakusa
Hotel Top Asakusa
Ryokan Shigetsu

Figure 3.4 Query composition

[10] We hypothesize that the data source provides a heuristic for the quality (e.g. timeliness or veracity) of data available from the source. Data that comes from an unknown database producer may benefit (or suffer) from integration and redistribution by compounding the positive (or negative) reputation of the integrator. If an unknown data source is cited in the Wall Street Journal, the perceived quality of the data might rise whereas if the data is cited in a daily tabloid known for exaggeration or hyperbole, the perceived quality of the data might fall.

Example 3.9 Attribution composition

Refer again to Q5 from example 3.8. We can translate the SQL queries into:

DRC5.1 {HNAME | (hotels(HNAME, ROOMS, PRICE) ∧ regions(HNAME, REGION) ∧
sites("Nakamise Dorsi", REGION)) ∨ (hostels(HNAME, PRICE, STATION) ∧
sites("Nakamise Dorsi", STATION))}

DRC5.2 {HNAME | (hotels(HNAME, ROOMS, PRICE) ∧ regions(HNAME, REGION) ∧
sites("Nakamise Dorsi", REGION)) ∨ (jyh(HNAME, PRICE, STATION, PHONE, FAX)
∧ sites("Nakamise Dorsi",STATION)) ∨ (rg(HNAME, PRICE, STATION, PHONE) ∧
sites("Nakamise Dorsi",STATION))}

Recall also that predicate hostels(XYZ) in DRC5.1 corresponds to

hostels(HNAME, PRICE, STATION, PHONE) $\overset{def}{=}$ jyh(HNAME, PRICE, STATION, PHONE,
FAX) ∨ rg(HNAME, PRICE, STATION, PHONE)

Regardless of how the query is posed, the result is the list of hotels and hostels:

Asakusa View, Ryokan Shigetsu, Sky Court Asakusa, and Hotel Top Asakusa

Step 1 in the algorithm is to collect the substitutions for the composed query, DRC 5.1. For brevity, we will only illustrate the composition of the relevant substitution. The relevant substitutions are:

{<hotels("Asakusa View"/HNAME); regions("Asakusa View"/HNAME, "Asakusa"/REGION);
sites("Nakamise Dorsi"/SNAME, "Asakusa"/REGION)>;
<hostels("Ryokan Shigetsu"/HNAME, "Asakusa"/STATION); sites("Nakamise Dorsi"/SNAME,
"Asakusa"/STATION)>;
<hostels("Sky Court Asakusa"/HNAME; "Asakusa"/STATION); sites("Nakamise
Dorsi"/SNAME, "Asakusa"/STATION)>;
<hostels("Hotel Top Asakusa"/HNAME) "Asakusa"/STATION); sites("Nakamise
Dorsi"/SNAME, "Asakusa"/STATION)>}

Informally, in Step 2 of the algorithm, we find the variables applicable to the composed relation, hostels and attribute those values against DRC 4.2. Yielding the following substitutions:

{<rg("Ryokan Shigetsu"/HNAME)>;
<rg("Sky Court Asakusa"/HNAME)>;
<rg("Hotel Top Asakusa"/HNAME)>;
<jyh("Sky Court Asakusa"/HNAME)>}

To complete the attribution composition, in Step 3, we combine the respective substitutions:

{<hotels("Asakusa View"/HNAME); regions("Asakusa View"/HNAME, "Asakusa"/REGION);
sites("Nakamise Dorsi"/SNAME, "Asakusa"/REGION)>;
<rg("Ryokan Shigetsu"/HNAME, "Asakusa"/STATION); sites("Nakamise Dorsi"/SNAME,
"Asakusa"/STATION)>;
<rg("Sky Court Asakusa"/HNAME; "Asakusa"/STATION); sites("Nakamise Dorsi"/SNAME,
"Asakusa"/STATION)>;
<rg("Hotel Top Asakusa"/HNAME) "Asakusa"/STATION); sites("Nakamise Dorsi"/SNAME,
"Asakusa"/STATION)>;

<jyh("Sky Court Asakusa"/HNAME; "Asakusa"/STATION); sites("Nakamise Dorsi"/SNAME, "Asakusa"/STATION)>}

This ultimately translates to the following relevant attribution:

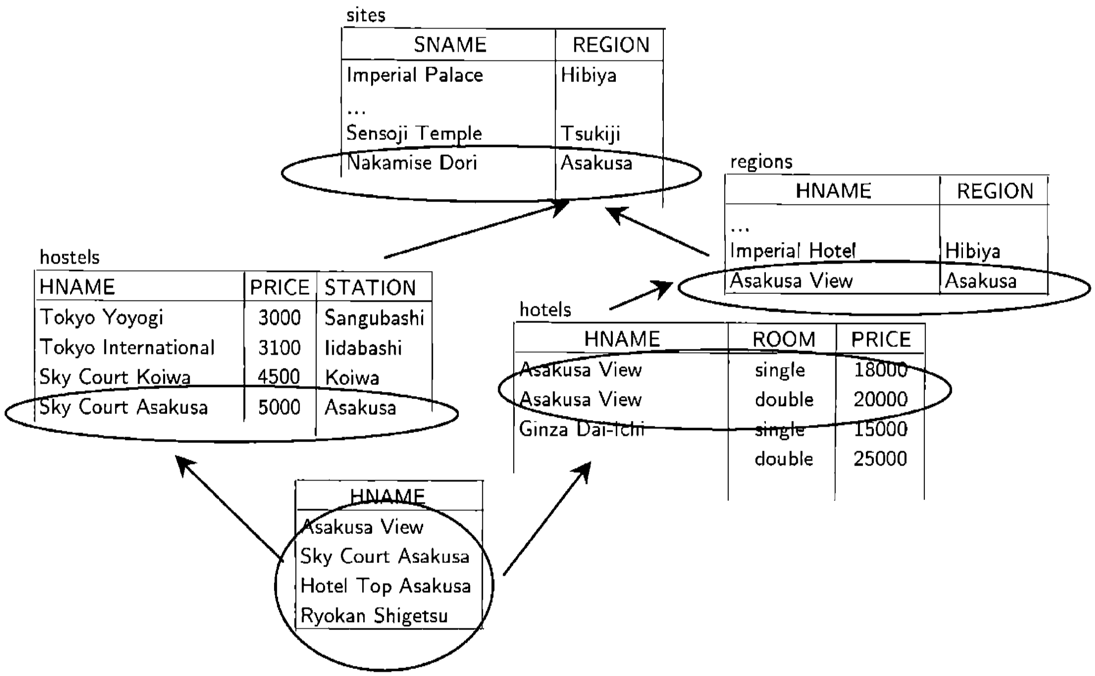
{<hotels; regions; sites>; <rg; sites>; <jyh; sites>}

The process of composing an attribution by iteratively tracing backwards through the constituent inputs is depicted in Figure 3.5.□

In looking more closely at Examples 3.6 and 3.9, we see that certain data values, such as the hostel "Sky Court Asakusa" may appear multiple times. This observation hints at a second category of multiple derivations, those within a single expression.

We originally separated multiple derivations into two categories: derivations from multiple expressions and derivations within a single expression. We can further separate derivations from a single expression into cases of weak equivalence and cases of natural join.

Weak equivalence encompasses the idea that tuples in a query result may differ only in their attribution. A straightforward example of this occurs in the case of relational union.



sites

SNAME	REGION
Imperial Palace	Hibiya
...	
Sensoji Temple	Tsukiji
Nakamise Dori	Asakusa

regions

HNAME	REGION
...	
Imperial Hotel	Hibiya
Asakusa View	Asakusa

hostels

HNAME	PRICE	STATION
Tokyo Yoyogi	3000	Sangubashi
Tokyo International	3100	Iidabashi
Sky Court Koiwa	4500	Koiwa
Sky Court Asakusa	5000	Asakusa

hotels

HNAME	ROOM	PRICE
Asakusa View	single	18000
Asakusa View	double	20000
Ginza Dai-Ichi	single	15000
	double	25000

HNAME
Asakusa View
Sky Court Asakusa
Hotel Top Asakusa
Ryokan Shigetsu

Figure 3.5 Attribution composition: Step 1

Example 3.10 Weak equivalence in union

```
SQL 4.1   select * from hostels

SQL 4.2   select HNAME, PRICE, STATION from jyh
          union
          select HNAME, PRICE, STATION from rg
```

in SQL 4.1, there is one substitution associated with Sky Court Asakusa

g("Sky Court Asakusa"/HNAME, 5000/PRICE, "Asakusa"/STATION, "81-3-3672-4411"/PHONE)

But in 4.2 there are TWO, one associated w/ querying rg (*r*) and one associated w/ querying jyh (*s*)

r("Sky Court Asakusa"/HNAME, 5000/PRICE, "Asakusa"/STATION, "81-3-3672-4411"/PHONE);

s("Sky Court Asakusa"/HNAME, 5000/PRICE, "Asakusa"/STATION, "81-3-3672-4411"/PHONE, "81-3-3875-4941"/FAX)} □

Similar behavior is exhibited when projecting a list of attributes that do not constitute a candidate key.

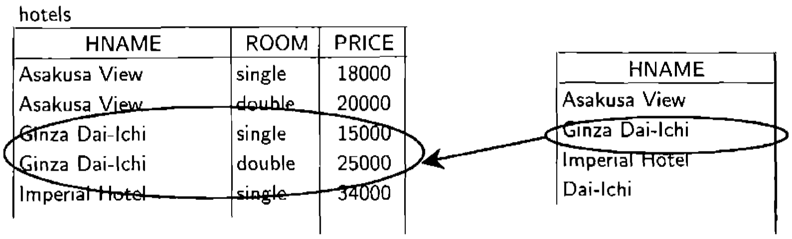
Example 3.11 Weak equivalence in projection

```
SQL 3.2   select HNAME from hotels
```

Pick one of the hotels in the result, for example. As seen in Figure 3.6, for each HNAME in the relation hotels, There are two lists of substitutions:

{("Ginza Dai-Ichi"/HNAME, "single"/ROOMS, 15000/PRICE)>;

<("Ginza Dai-Ichi"/HNAME, "double"/ROOMS, 25000/PRICE)>} □



hotels

HNAME	ROOM	PRICE
Asakusa View	single	18000
Asakusa View	double	20000
Ginza Dai-Ichi	single	15000
Ginza Dai-Ichi	double	25000
Imperial Hotel	single	34000

HNAME
Asakusa View
Ginza Dai-Ichi
Imperial Hotel
Dai-Ichi

Figure 3.6 Weak duplicates

Though logical models of relations, like the relational calculus, rely upon set semantics, this theory of attribution treats every instance of a tuple as unique and having a distinct attribution with respect to the query and underlying data sources. To preserve the set semantics of the relational data model, attributions for weak duplicates are combined together.

The second category of duplication, that occurs in single expressions, stems from looking for relationships between relations (called a join operation) rather than taking the union of different relations. Informally, we want to distinguish between comparisons on values that represent the same thing and values that merely "look" alike.[11] We call values that represent the same thing duplicates. However, we would like to treat values that merely "look" alike somewhat differently.

Example 3.12 Multiple derivations in joins.
To explore this issue, we will reconsider Q5 from earlier. However, this time, we separate the query explicitly into:
Q6. Identify hotels around the Nakamise Dori and
Q7. Identify hostels around the Nakamise Dori.
These queries translate to SQL 6 and SQL 7, as indicated below. Separating the queries this way will allow us to look more carefully at how values are compared between tables.

```
SQL6    select HNAME
        from hotels, regions, sites
        where hotels.HNAME = regions.HNAME
        and regions.REGION = sites.REGION
        and sites.SNAME = "Nakamise Dorsi"

SQL7    select HNAME
        from hostels, sites
        where hostels.STATION = sites.REGION
        and sites.SNAME = "Nakamise Dorsi"
```

We translate the above SQL queries into the following DRC expressions:

DRC 6 {HNAME | hotels(HNAME, ROOMS, PRICE) ∧ regions(HNAME, REGION) ∧ sites("Nakamise Dorsi", REGION)}

DRC 7.1 {HNAME | hostels(HNAME, PRICE, STATION) ∧ sites("Nakamise Dorsi", STATION)}

DRC 7.2 {HNAME | hostels(HNAME, PRICE, STATION) ∧ sites("Nakamise Dorsi", REGION) ∧ (STATION = REGION)}

(where DRC 6 and DRC 7.1 are the subformulas that we used in DRC 5 and DRC 7.2 is a logically equivalent expression to DRC 7.1

To find hotels around Nakamise Dorsi, we use geographic region names associated with tourist attractions and also associated with the hotel addresses. Unfortunately, we do not have such information available for the youth hostels. Instead, we match the regions for the local tourist attractions with the names of railroad stations. This is illustrated in Figure 3.7. The scalar values are the same, but they come from different domains. This distinction is made explicit in the calculus by the distinction between multiple occurrences of the same variable

[11] We consider natural join as distinct from theta comparison where theta is equality. Natural join is represented in the relational calculus as multiple occurrences of the same variable in two or more predicates. In the (named) relational algebra, it corresponds to the idea that different relations may include the same domain. Using a slight variation on the standard notation, this is represented by identical attribute names in multiple relation schemes.

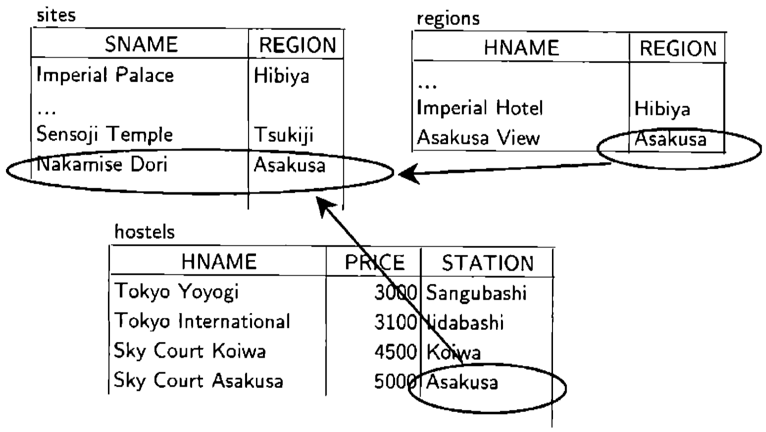
versus explicit equality. Consider a few of the comprehensive substitutions for the expressions from Example 3.12.

comprehensive substitution for DRC 6
<"Asakusa View"/HNAME, "single"/ROOMS, 18000/PRICE, "Nakamise Dorsi"/SNAME,
"Asakusa"/REGION>

comprehensive substitution for DRC 7.1
<"Ryokan Shigetsu"/HNAME, 7000/PRICE, "Asakusa"/STATION, "Nakamise Dorsi"/SNAME>

comprehensive substitutions for DRC 7.2
<"Ryokan Shigetsu"/HNAME, 7000/PRICE, "Asakusa"/STATION, "Nakamise Dorsi"/SNAME,
"Asakusa"/REGION>

The intuition is that multiple occurrences of the same variable constitute a *single substitution* that derives from multiple sources. The substitution "Asakusa"/REGION in DRC 6 stems from two distinct sources; hotels and region. Explicit equality, by contrast, suggests that the equated variables are different values with their own substitution. The substitution "Asakusa"/REGION is equated with the substitution "Asakusa"/STATION in DRC 7, but we do not consider these substitutions to share the same sources. The relation hostels is not a source for "Asakusa"/REGION even though the variables are equated and the relation hotels is considered a source.[12] □



sites

SNAME	REGION
Imperial Palace	Hibiya
...	
Sensoji Temple	Tsukiji
Nakamise Dori	Asakusa

regions

HNAME	REGION
...	
Imperial Hotel	Hibiya
Asakusa View	Asakusa

hostels

HNAME	PRICE	STATION
Tokyo Yoyogi	3000	Sangubashi
Tokyo International	3100	Iidabashi
Sky Court Koiwa	4500	Koiwa
Sky Court Asakusa	5000	Asakusa

Figure 3.7 Multiple derivations in joins

Note the implicit equivalence between multiple occurrences of the same variable versus the explicit built-in theta-comparison predicate (X=Y) in calculus expressions. We arrive at this conclusion by substituting all occurrences of Y with X and eliminating the explicit theta-

[12] Note that railroad station names and geographic region names may not always coincide. The example here is intended to illustrate situations where values from different domains are used in comparisons and query conditions, suggesting distinct lineage.

comparison. Because our intuition for attribution makes use of the distinction despite the implicit equivalence, we conclude that the different types of attribution are not equivalent for equivalent expressions when we allow built-in predicates for explicit equality.

Negation is the other place in which our observations on the attribution of equivalent queries breaks down. Our intuition is that attribution corresponds to those substitutions that correspond to a true interpretation. What then is the substitution that makes a statement about the non-existence of something true? Applying the conventional database interpretation of negation, we suggest that the way to prove a negative substitution is by comparing that substitution to all known positive substitutions. If the item of interest is not known to be true, we conclude that it must be false.[13]

Example 3.13 Negation

We take our original query and invert it.

Q8. Hotels NOT by the Imperial Palace. We can write this in SQL as:

```
SQL 8.1    select HNAME
           from regions
           where HNAME not in (
                   select HNAME
                   from regions, sites
                   where regions.REGION = sites.REGION
                   and sites.SNAME = "Imperial Palace")
```

One possible interpretation of this expression in the DRC is:

DRC 8.1 {HNAME | regions(HNAME, REGION) ∧ ¬(regions(HNAME, REGION) ∧
sites("Imperial Palace", REGION)}

The answer to DRC 8.1 is: Hotel Sofitel, Katsutaro, Asakusa View

We know that the Asakusa View is not by the Imperial Palace. What are the corresponding substitutions into DRC8.1 indicating the truth of this? We need to establish, in a positive sense, what hotels <u>are</u> by the Imperial Palace and then, given a fixed list of hotels, (those in regions), we keep the remainder. This process is depicted in Figure 3.8.

Consider, for brevity, just one comprehensive attribution for DRC 8.1:

<*f*("Asakusa View"/HNAME, "Asakusa"/REGION)
g("Dai-Ichi"/HNAME, "Hibiya"/REGION, "Imperial Palace"/SNAME)
g("Imperial"/HNAME, "Hibiya"/REGION, "Imperial Palace"/SNAME)> □

To see some of the difficulty created by negation, consider an equivalent expression to DRC 8.1, which we present in Example 3.14.

[13] The negation as failure interpretation adopted in the database community suggests that a negated subformula is true only when no true interpretation of the subformula is found (Ullman 1988).

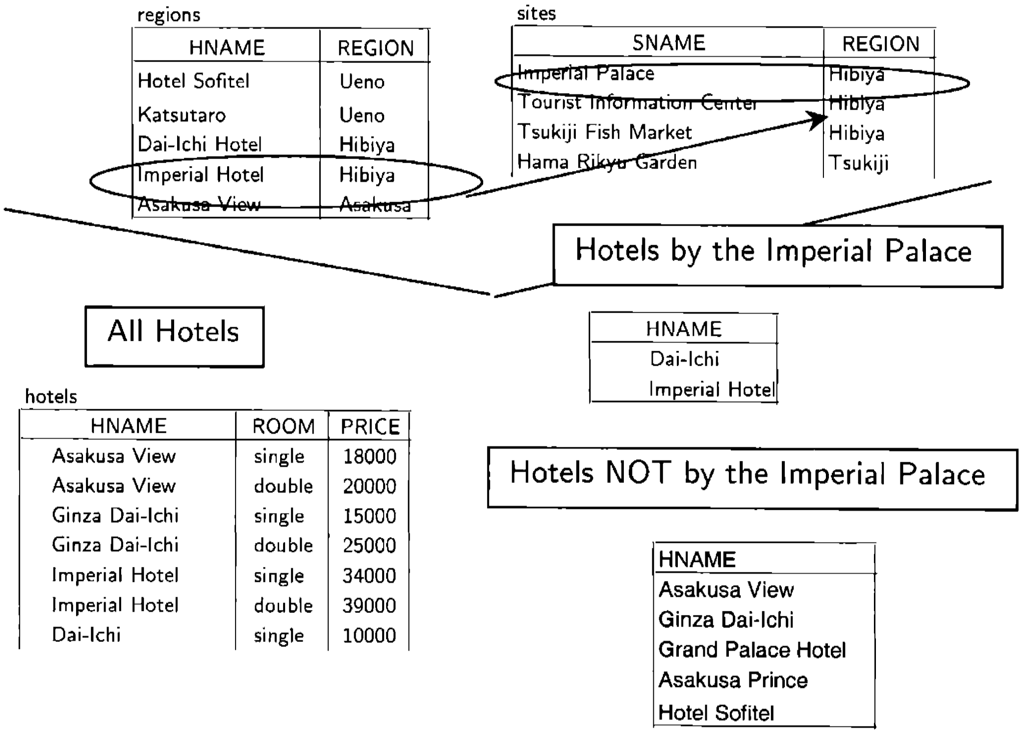
Example 3.14 Attribution equivalence breaks down under negation

DRC 8.2 {HNAME | regions(HNAME, REGION) ∧ ¬sites("Imperial Palace", REGION)}

DRC 8.2 is logically equivalent to DRC 8.1. The difference is that we have pushed the negation down to the atoms and then distributed the conjuncts and disjuncts.

Compare the comprehensive substitution associated with the hotel Asakusa View that we examined in Example 3.13. Notice

<*f*("Asakusa View"/HNAME, "Asakusa"/REGION)
g("Imperial Palace"/SNAME, "Hibiya"/REGION)> □



regions

HNAME	REGION
Hotel Sofitel	Ueno
Katsutaro	Ueno
Dai-Ichi Hotel	Hibiya
Imperial Hotel	Hibiya
Asakusa View	Asakusa

sites

SNAME	REGION
Imperial Palace	Hibiya
Tourist Information Center	Hibiya
Tsukiji Fish Market	Hibiya
Hama Rikyu Garden	Tsukiji

Hotels by the Imperial Palace

HNAME
Dai-Ichi
Imperial Hotel

All Hotels

hotels

HNAME	ROOM	PRICE
Asakusa View	single	18000
Asakusa View	double	20000
Ginza Dai-Ichi	single	15000
Ginza Dai-Ichi	double	25000
Imperial Hotel	single	34000
Imperial Hotel	double	39000
Dai-Ichi	single	10000

Hotels NOT by the Imperial Palace

HNAME
Asakusa View
Ginza Dai-Ichi
Grand Palace Hotel
Asakusa Prince
Hotel Sofitel

Figure 3.8 Negation in attribution

3.3 Levels of attribution

In our last two Examples 3.13 and 3.14 we examined only a subset of the substitutions. In particular, we considered only those substitutions corresponding to the result that the hotel Asakusa View is not by the Imperial Palace. This suggests that, rather than speaking about the attribution for a query result, we might wish to consider attributing only one part of the result. Returning to the analogy of a bibliography, perhaps the reader is only interested in Part 1 of the thesis. As noted in the Introduction, Part 1 consists of Chapters 3, 4, 5, and 6. Taken together, these four chapters present our model of attribution. However, our bibliography, which follows Chapter 9, is a single list of all works referenced throughout the entire document. The reader might only wish to know the works which were referenced in Chapter 3 – 6. Perhaps the reader is only interested in Chapter 3.

And because we know that the result of one query can become the source for another query (query composition), we extend the idea of attributing one part of the result to the idea that we might attribute with only part of a source rather than attribute using the relation as a whole. We refer to these ideas as result and source granularity respectively.

The general intuition is that attribution defines pointers or references from a query result back to its constituent sources. Granularity therefore corresponds to the pointer's precision. Beginning with result granularity, at the finest granularity, we might wish to attribute a specific instance of a value in a result. More generally, we might think of all instances of a value in a result. Further coarsening our granularity, we could consider the attribution corresponding to a range of values (e.g. an entire tuple, a set of tuples, or perhaps a column). At the limit, we could attribute the entire result relation.

Example 3.15 Result granularity

Consider Q9, Hotel names, hotel prices, and names of sites around Tokyo, Japan.

SQL 9 select HNAME, PRICE, SNAME
from hotels, sites

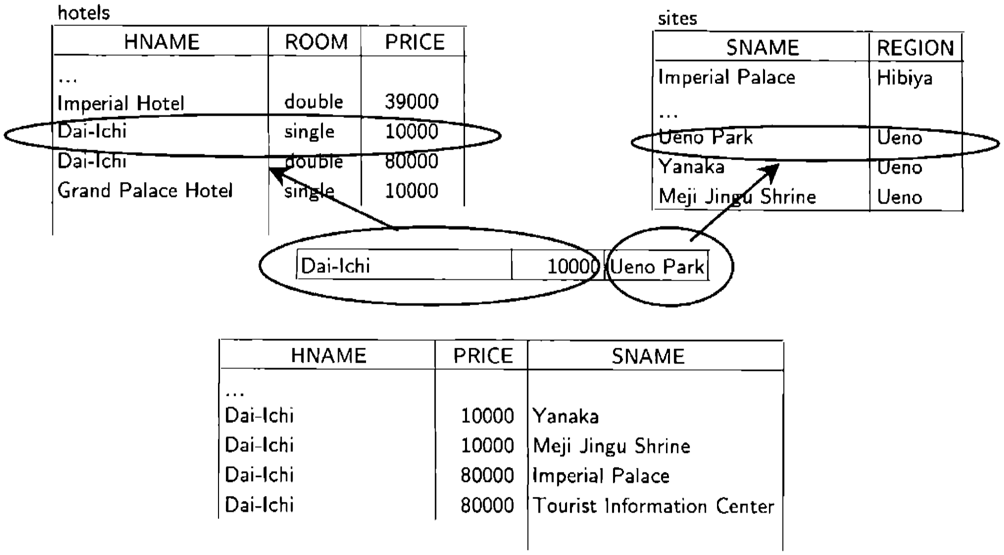
DRC 9.1 {HNAME, PRICE, SNAME | hotels(HNAME, ROOM, PRICE) ∧ sites(SNAME, REGION)}

As seen in Figure 3.9, we can discuss the attribution associated with the specific instance of a result where HNAME = "Dai-Ichi" (corresponding to one tuple). Single rooms by Ueno Park correspond to the following tuple: <"Dai-Ichi", 10000, "Ueno Park">

The corresponding comprehensive attribution is:
{<*f*("Dai-Ichi"/HNAME, "single"/ROOM, 10000/PRICE, "Ueno Park"/SNAME, "Ueno"/REGION)>}

We could ask for all instances of "Dai-Ichi" hotel in the result. Because the query is a Cartesian product, the actual solution is quite large, but one part of it includes Table 3.6 with the following substitutions:

{... < *f*("Dai-Ichi"/HNAME, "single"/ROOM, 10000/PRICE, "Yanaka"/SNAME, "Ueno"/REGION)>;
f("Dai-Ichi"/HNAME, "single"/ROOM, 10000/PRICE, "Meji Jingu Shrine"/SNAME, "Ueno"/REGION)>;
f("Dai-Ichi"/HNAME, "double"/ROOM, 80000/PRICE, "Imperial Palace"/SNAME, "Hibiya"/REGION)>;
f("Dai-Ichi"/HNAME, "double"/ROOM, 80000/PRICE, "Tourist Information Center"/SNAME, "Hibiya"/REGION)>; ...} □



hotels

HNAME	ROOM	PRICE
...		
Imperial Hotel	double	39000
Dai-Ichi	single	10000
Dai-Ichi	double	80000
Grand Palace Hotel	single	10000

sites

SNAME	REGION
Imperial Palace	Hibiya
...	
Ueno Park	Ueno
Yanaka	Ueno
Meji Jingu Shrine	Ueno

Dai-Ichi	10000	Ueno Park

HNAME	PRICE	SNAME
...		
Dai-Ichi	10000	Yanaka
Dai-Ichi	10000	Meji Jingu Shrine
Dai-Ichi	80000	Imperial Palace
Dai-Ichi	80000	Tourist Information Center

Figure 3.9 Source/result granularity

Dai-Ichi	10000	Yanaka
Dai-Ichi	10000	Meji Jingu Shrine
Dai-Ichi	80000	Imperial Palace
Dai-Ichi	80000	Tourist Information Center

Table 3.6 Result granularity

Likewise, we might draw the parallel conclusions for source granularity. We could attribute using a specific instance of a substitution (e.g. the source tuple that corresponds to the specific instance of a substitution), all occurrences of a substitution in a particular source (e.g. every tuple in a source that provides the substitution), or again at the extreme, the name of the relation that corresponds to true substitutions.

Example 3.16 Source granularity

Throughout the Chapter, we have given answers for sources as relation names. Using DRC 9 from Example 3.15, however, we can provide references to the sources with varying levels of precision as well.

Attribution for DRC 9 as sources:

<hotels; sites>

We can also give:

hotels("Dai-Ichi"/HNAME)

which implicitly indicates all instances in the hotels relation where HNAME = "Dai-Ichi"
{<"Dai-Ichi", "single", 10000>; <"Dai-Ichi", "double", 80000>}

or we can give an explicit instance of "Dai-Ichi" in the source relation
<hotels("Dai-Ichi"/HNAME, "single"/ROOM, 10000/PRICE)> □

Given that we expressed our intuition about attribution in terms of an expression and a result, how might we express interest in the attribution for an explicit granule rather than the relation as a whole, given that we have been thinking about attribution in terms of answers to queries? Conceptually, we know that we can think of substitutions that make a particular substitution for the free variables (one tuple in the result) true. However, within our framework of attribution for relations, we might also take our cue from the observation that the relational calculus is closed. Closure permits us, as demonstrated earlier, to compose queries. At the same time, we know that we can compose attribution as well. Consequently, if we want all instances or specific instances of values in the result, we propose composing a query on the result and then composing the corresponding attribution to return the attribution for the result granule of interest.

Example 3.17 Specifying granularity
We refer again to Q9 Hotel names, hotel prices, and names of sites around Tokyo, Japan.
SQL 9 select HNAME, PRICE, SNAME
from hotels, sites

Intuitively, if we are interested in a result granule defined as, all instances of "Dai-Ichi" in the result, we think of something like:
DRC 9.2 {"Dai-Ichi", PRICE, SNAME | hotels(HNAME, ROOM, PRICE) $\wedge$ sites(SNAME, REGION)}

In other words, we want all substitutions in the answer where the HNAME is "Dai-Ichi." We can construct just such a query if we think of:
DRC 10 {HNAME, ROOM, PRICE, SNAME | temp("Dai-Ichi", ROOM, PRICE, SNAME) $\wedge$ (HNAME = "Dai-Ichi")}

Where temp(HNAME, ROOM, PRICE, SNAME) $\overset{\text{def}}{=}$ {HNAME, ROOM, PRICE, SNAME | hotels(HNAME, ROOM, PRICE) $\wedge$ sites(SNAME, REGION)}

We might also think of a subset of instances of "Dai-Ichi" in the result. Consider:
DRC 11 {HNAME, ROOM, PRICE, SNAME | temp(HNAME, ROOM, PRICE, SNAME) $\wedge$ (HNAME = "Dai-Ichi") $\wedge$ (ROOM = "single") $\wedge$ (PRICE = 10000)} □

Regardless of source granularity, the comprehensive attribution for a value in result tuple is the same for every other value in the same tuple. This makes sense. A DRC expression corresponds to a set of tuples. Therefore, one list of substitutions that makes one instance of the expression true applies to every value in the corresponding result tuple. Likewise, given relation-level source granularity, the comprehensive attribution for every value in the result is the same. Again this makes intuitive sense. This merely articulates the observation that all of the relations in the WHERE clause of an SQL statement apply to the relation as a whole.

Note by contrast that for source or relevant attribution, the attribution of different values or tuples are not necessarily the same. In the UNION case, we saw how weak duplicates illustrated a single tuple might have more than one source. As a more subtle case, refer again to the Cartesian product of Q9. From Figure 3.9, we see how distinct sources can associate with only a subset of attributes in a result relation.

In summary, we list the different features and properties captured by our model of attribution and discussed throughout this Chapter. We present the model more formally in Chapter 4 and subsequently propose an extension to the relational algebra to operationalize one instantiation of our theory. In particular, we demonstrate attribution using relation-level source granularity. We conclude Part 1 by considering how the theory might extend into the semi-structured environment of the Web.

- Attribution refers to the substitutions that make the interpretation of the expression true.
- In the case of negation, we use negation as failure semantics to establish that a predicate does not hold.
- There are three distinct types of attribution: comprehensive, source, and relevant.
- There are a number of ways in which a query result might have more than one attribution:
    - Multiple queries for the same result
    - Weak equivalence
    - Strict equivalence (equivalent expressions using only base relations)
    - Equivalence using composed data sources
    - Weak duplicates
    - Multiple instances of the same variable in an expression (e.g. natural join)
- For conjunctive queries with theta-comparisons but omitting explicit equality, the comprehensive and source attribution of equivalent queries is equivalent.
- For positive queries, the relevant attribution of equivalent queries is equivalent
- We can compose the attribution of composed queries (where there is no more than one level of negation) by recursively unfolding and attributing sub-queries in a depth-first manner.
- Weak duplicates and multiple occurrences of the same value in different predicates of a calculus expression (join variables in a natural join) entail multiple derivations of the same row or column (tuple or attribute domain).
- Theta comparisons involving explicit equality represent different values, each with their own, distinct derivation.
- We can attribute using different levels of granularity on the source side and attribute different result granules.

- We specify the attribution of different result granules by composing queries on the original result.

4 Formal model

In this Chapter, we present our formal model of attribution along with a number of properties of the model. Section 1 offers a brief overview of the domain relational calculus for those unfamiliar with the formalism. Section 2 introduces our definition of attribution in the context of the syntax of a domain relational calculus expression. To formalize the model, we begin with the set of conjunctive queries (defined below) and gradually expand the query language in the standard way.[14] We conclude the chapter by relating the formal model back to the desiderata originally specified in Chapter Two.

4.1 The domain relational calculus

Our formalization of attribution is based upon the Domain Relational Calculus (DRC). For those already familiar with the DRC, we begin by specifying the calculus syntax and notation used in the remainder of this thesis. For those unfamiliar with the DRC, we follow our specification with a brief overview. The DRC is built upon, and our overview assumes, basic familiarity with the first-order predicate calculus.

4.1.1 Syntax and notation

We use the set of lists notation for a relation. Following (Ullman 1988; 1989), at times, we make selective use of variable names to denote attributes for readability. We define attribution in terms of the interpretation (Maier 1983) of a safe DRC expression, where safety is defined syntactically by (Ullman 1988).

A list of substitutions $a = <c_1/X_1, c_2/X_2, \ldots, c_n/X_n>$ projected on a formula f, written $a(f)$, returns the sub-list of substitutions for the variables in f. A list of substitutions a is in the attribution for an expression $E = \{x \mid f(x)\}$ when a has the minimal number of substitutions required to recursively interpret every sub-formula f' of f such that $s(f) = c$ and $I(f(c/x)) = true$.

Furthermore, all expressions are assumed to be in Safe Range Normal Form (SRNF) and Relational Algebra Normal Form (RANF) meaning negations are pushed down to atoms and existential quantification and connectives are flattened (Abiteboul, Hull, and Vianu 1995).

[14] See (Ullman 1988) and (Abiteboul, Hull, and Vianu 1995).

We further assume, consistent with RANF, that formulas without negation are expanded into prenex disjunctive normal form (DNF). Given the syntactic safety rules, each disjunct therefore projects all and only the set of free variables in the expression.

As a shorthand, for expressions of the form:
$\{X_1,X_2,\ldots,X_n \mid \exists Y_1, \exists Y_2,\ldots,\exists Y_m)\ f(X_1,X_2,\ldots,X_n,Y_1,Y_2,\ldots,Y_m)\}$
We will sometimes substitute:
$\{X_1,X_2,\ldots,X_n \mid (\exists Y_1,Y_2,\ldots,Y_m)\ f(X_1,X_2,\ldots,X_n,Y_1,Y_2,\ldots,Y_m)\}$
And when obvious, we will omit the existential quantification entirely:
$\{X_1,X_2,\ldots,X_n \mid f(X_1,X_2,\ldots,X_n,Y_1,Y_2,\ldots,Y_m)\}$

Following (Ullman 1988), we use Extensional Data Base (EDB) to refer to base relations and Intentional Data Base (IDB) to refer to relations composed on base relations (e.g. views).

4.1.2 A review of the calculus

Let **D** be a set of disjoint domains over which all relations are defined. A relational scheme is a pair (J,D) where J is an index (a set of integers from 1 to $max(J)$) and D is a function, $D: J \rightarrow \mathbf{D}$. A relation is then defined over a scheme as a finite subset of the Cartesian Product of the domains in the scheme. A tuple is therefore a list of values where the J^{th} value is drawn from the corresponding domain and a relation is a finite set of such lists.

In practice, the set-of-lists notation is equivalent to more conventional attribute-value naming (Ullman 1988). Following Ullman and (Abiteboul, Hull, and Vianu 1995), where obvious to do so, we may use carefully selected variable names to denote particular attribute domains. We may then denote a relation scheme by a tuple instance consisting entirely of domain variables, an ordered list of variable names $(A_1,A_2, \ldots ,A_{max(J)})$ where each A_J is a variable name for a value drawn from $D(J)$.

Harkening back to our motivating example from Chapter 3, variable names might include: NAME, PRICE, REGION, ROOM, STATION, etc.

Definition 4.1 Atomic formulas
Basic formulas in the domain calculus (also called *atomic formulas*) are expressed in terms of relations, domain variables, and Θ, the set of comparison operators (e.g. >, ≥ , =, ≤, <) for every domain in **D**.

1. If r is a relation in d with scheme $(A_1,A_2, \ldots ,A_n)$ then $r(X_1,X_2,\ldots,X_n)$ is an atomic formula where X_i is either a domain variable for D_i (e.g. of type D_i) or a constant $c_i \in D_i$.
2. If X and Y are domain variables and c is a constant drawn from the appropriate domain, then $X\ \theta Y$, $X\ \theta c$, and $c\ \theta X$ are all atomic formulae.
3. The Boolean constants *true* and *false* are also atomic formulae. □

Example 4.1 Atomic formulas
hotels(HNAME,ROOM,PRICE) is a predicate for the relation hotels
hotels('Asakusa View', 'single', 18000) is an atomic formula
hotels (HNAME, 'single', PRICE) is also an atomic formula as are
hotels.HNAME = regions.HNAME and PRICE < 90,000. □

Definition 4.2 Calculus formula
We recursively extend our definition of a calculus formula by building upon our atomic formulae using the logical connectives ($\neg$, $\wedge$, $\vee$) and the quantifiers ($\exists$, $\forall$) in a manner similar to the predicate calculus.

1. If f is a formula, then $\neg f$ is a formula.
2. If f and g are both formulas, then $f \wedge g$ is a formula as is $f \vee g$.
3. If f is a formula and X is a domain variable, then $\exists X f$ and $\forall X f$ are both formulas where free occurrences of X in f are bound by $\exists X$ and $\forall X$ respectively using the expected definitions for free and bound (Maier 1983 at 231; Ullman 1988 at 147).
4. If f is a formula, then (f) is a formula

The parentheses explicitly define groupings of operands as we might expect. In the absence of parentheses, the quantifiers $\exists X$, and $\forall X$ have highest, equal precedence. $\neg$, $\wedge$, $\vee$ follow in decreasing order of precedence. □

Example 4.2 Calculus formulas
If regions(HNAME, Hibya) is a formula, then ¬ regions(HNAME, Hibya) is a formula.
It therefore follows that (¬ regions(HNAME, 'Hibya')) is a formula.
(¬ regions(HNAME, 'Hibya')) ∧ hotels(HNAME,ROOM,PRICE) is also a formula.
Using the quantifiers in conjunction with parentheses can result in some subtly different formulas.
∃HNAME(¬ regions(HNAME, 'Hibya')) ∧ hotels(HNAME,ROOM,PRICE) is not equal to
∃HNAME((¬ regions(HNAME, 'Hibya')) ∧ hotels(HNAME,ROOM,PRICE)). □

We offer a brief aside on the legality of formulas and note that domain variables should be used consistently so that in the formula ∃X((¬ regions(X, 'Hibya')) ∧ hotels(X,Y,Z)), domain variable X refers to the domain of lodging establishment names and the formula ∃X((¬ regions(X, 'Hibya')) ∧ hotels(Z, X, X)) is somewhat nonsensical (Maier 1983 at 231).

Given a formula f, we would like to know what that formula means. Following Maier, we first define a *substitution*. We then arrive at an *interpretation* of f based upon a substitution for the free variables in f and the expected meaning of the logical connectives and quantifiers. The following definitions for substitutions and interpretations are the foundation of our formalism for attribution.

The intuition behind the substitution is to recall that formulas are defined with respect to a set of base relations called a database *d*. Atomic formulas for relation *r* in database *d* correspond to base tables in *d* (or constraints that take the form of comparisons on values that appear in one or more initial tables.) A substitution is a "random" replacement of all free variables in a formula with constants from their corresponding attribute domains. An atomic formula denoted by $r(X_1,X_2,\ldots,X_n)$ is *true* for all and only the substitutions $(c_1,c_2,\ldots,c_n)$ that are in the base table $r \in d$.

Definition 4.3 Substitution
More formally, let $f(X_1,X_2,\ldots,X_n)$ be a legal calculus formula as defined earlier where $X_1,X_2,\ldots,X_n$ corresponding to their respective domains are the only free domain variables in *f*.

A *substitution* of a tuple $(c_1,c_2,\ldots,c_n)$ in $f(X_1,X_2,\ldots,X_n)$ is denoted by $f(c_1/X_1,c_2/X_2,\ldots,\ c_n/X_n)$ where $c_i \in D_i$, the domain corresponding to A_i. We rewrite *f*, replacing every free occurrence of X_i with c_i. Ground atoms, atomic formulae containing only constants ($\underline{k_i}$) following the substitution, are replaced with *true* or *false* as follows:

1. If the ground atom is a relation $r(k_1,k_2,\ldots,k_m)$ then replace the atom in the formula *f* with *true* if tuple $(k_1,k_2,\ldots,k_m) \in r$.
2. If the ground atom is a comparison $k_i\ \theta\ k_j$ then replace the atom in the formula *f* with *true* or *false* as appropriate. □

Example 4.3 Substitution
Consider the following formulas based upon the travel database of Chapter 3:
f = sites(SNAME, REGION)
g = ∃ADDRESS (hr(HNAME, REGION, ADDRESS) ∧ sites(SNAME, REGION))

Suppose that the domain of tourist attractions included :
{'Imperial Palace', 'Yanaka', 'Fanueil Hall', 'Revere House', 'Tower of London'}

and that the domain of regions included:
{'North End', 'Beacon Hill', 'Hibiya', 'Asakusa', 'Ueno'}

then the following substitutions would be:
f('Imperial Palace'/SNAME, 'Beacon Hill'/REGION) = false
f('Revere House'/SNAME, 'North End'/REGION) = false[15]
g('Dai-Ichi'/HNAME, 'Yanaka'/SNAME, 'Ueno'/REGION) =
∃ADDRESS (hr('Dai-Ichi', 'Ueno', ADDRESS) ∧ *true*). □

[15] note that the Revere House may indeed be in the North End, but this is not a fact in the relation *sites*. Therefore the predicate evaluates to *false*.

FORMAL MODEL

Definition 4.4 Interpretation

The interpretation of a formula f with no free domain variables, written $I(f)$, is recursively defined as:

1. If f is *true* or *false* then $I(f)$ is *true* or *false*.
2. If f is $\neg\, g$ and g has no free variables, we say if $I(g)$ is *true*, $I(f)$ is *false*. Otherwise, $I(f)$ is *false*.
3. If f is $g \wedge h$ then $I(f)$ is *true* when both $I(g)$ and $I(h)$ are *true* and *false* otherwise.
4. If f is $g \vee h$ then $I(f)$ is *false* when both $I(g)$ and $I(h)$ are *false* and *true* otherwise.
5. If f is $\exists X(A)g$ where only X is free in g, then $I(f)$ is true when there is at least one value $c_i \in$ dom(A) for which $I(g(c/X))$ = true.
6. If f is $\forall X(g)$ where only X is free in g, then $I(f)$ is true when for every value $c_i \in$ dom(A) for which $I(g(c/X))$ = true.
7. If f is (g) then $I(f) = I(g)$. □

Definition 4.5 Domain relational calculus (DRC) expression

A calculus *expression* has the form $\{X_1,X_2,\ldots,X_n \mid f(X_1,X_2,\ldots,X_n)\}$ where, as indicated above, $f(X_1,X_2,\ldots,X_n)$ is a legal calculus formula and $X_1,X_2,\ldots,X_n$ corresponding to attributes $A_1,A_2, \ldots ,A_n$ are the only free domain variables in f. The value of an expression E on database d is therefore a relation r having scheme (J,D) for tuples of the form $(A_1,A_2, \ldots ,A_n)$ and containing all tuples $(c_1,c_2,\ldots,c_n)$ where $c_i \in D_i$ and $I(f(c_1/X_1,c_2/X_2,\ldots, c_n/X_n))$ = *true*. □

Example 4.4 A query as a domain relational calculus expression

We can translate the query which regions have a station or tourist attraction? Into the following expression:

{REGION | ∃SNAME, STATION (sites(SNAME, REGION) ∨ trains(STATION, REGION))} □

A domain calculus expression is therefore merely one way of articulating a query over a set of base relations. The expression $\{X_1,X_2,\ldots,X_n \mid f(X_1,X_2,\ldots,X_n)\}$ is a query for all tuples in the database that satisfy the query constraints in f. The answer to the query is a relation whose schema is $(A_1,A_2, \ldots ,A_n)$.

There is one significant problem with this definition of expressions and interpretations. Relations are defined as finite subsets of the Cartesian product of the domains $D_1 \times D_2 \times \ldots \times D_n$. However, domains themselves could be infinite.[16] The problem arises when we attempt to find an interpretation for legal calculus expressions that query infinite domains, possibly producing infinite relations.

[16] Consider the domain for the attribute price from our earlier examples. We certainly would not want the database to set an arbitrary bound on the maximum price a hotel could charge for one night's stay. Likewise, the domain for the attribute name might include hotels from around the world including "Le Meridien, Boston" in Boston, MA and the "Warwick Hotel" in Philadelphia, PA. The Cartesian product of name and price includes all possible permutations of hotels worldwide and an infinte range of prices. However, as indicated in our example database from Chapter 3, the relation hotels contains a finite subset of hotel names corresponding to establishments in Tokyo, Japan and only the corresponding prices charged by those Tokyo hotels.

Example 4.5 Negation and infinite relations[17]

In Q8 of Chapter 3, we asked, "List all hotels that are not in the same region as the *Imperial Palace* in Tokyo, Japan." Knowing that the Imperial Palace is in the Hibiya region of Tokyo, consider a simpler variant on this query which asks "List all hotels that are not in the Hibiya region of Tokyo." We might translate this query into the following calculus expression:

$\{\text{HNAME} \mid \neg\text{regions(HNAME,'Hibiya')}\}$.[18]

The answer, of course, is the presumably rather large set of substitutions for HNAME, (c/name) such that (c, 'Hibiya') is not a pair in table regions. □

To address the problem of infinite relations, we reach beyond the predicate calculus framework to further restrict the types of expressions that we consider evaluable. This concept is called *safety*.[19] The general intuition behind safe calculus expressions is that interpretation of domain variables somehow be explicitly constrained to some finite set of values. To guarantee this, we claim that value(s) which make the expression true must come from a domain consisting of all values that appear either in the constants or in the (finite) relations mentioned in the query.[20]

Example 4.6 Limited expressions

Rewrite the previous example to "bind" to a finite domain. In this case, we use Hibiya:
$\{\text{NAME} \mid \exists\text{REGION}(\text{regions(NAME, REGION)} \wedge \neg(\text{REGION} = \text{'Hibiya'}))\}$

Both the quantified variable region and the free variable name are limited by relation regions. In evaluating the existential quantifier, although the domain for variable region might be infinite, we need only consider values that appear in regions. Likewise, we only consider possible names that appear in regions. □

Definition 4.6 Safe DRC

Formally, we define the construction of a *safe* DRC formula following Ullman (1988; 1989) as one where every free variable must appear in at least one non-negated atomic formula that corresponds to a finite relation.

1. There are no uses of universal quantification.
2. Disjuncts must have the same set of free variables.

[17] We use the example of negation here, but similar problems exist for interpreting existential and universal quantification. See [Maier, 1983 #16 at 244-49.

[18] $\{\text{name} \mid \neg\text{regions(name,"Hibiya")}\}$ is shorthand. The formal expression would be $\{\text{name} \mid \exists\ \text{region}\ (\neg(\text{regions(name,region)}) \wedge (\text{region} = \text{"Hibiya"}))\}$. Because "Hibiya" is the only possible substitution for region, we remove the existential quantifier and substitute "Hibiya" as a constant in the remaining expression.

[19] There is a more general notion of safety that does not contribute to our definition of attribution and so is overlooked. See "limited evaluation" in [Maier, 1983 #16]or "domain independence" in (Ullman 1988).

[20] (Maier 1983)

3. For a maximal subformula that is the conjunction of one or more formulas $F_1 \wedge F_2 \wedge \ldots \wedge F_m$, variables free in any F_i must be *limited* such that
    3.1. A variable is limited if it is free in a formula F_i that is not a comparison and is not negated.
    3.2. If F_i is a comparison $X=c$ then X is limited.
    3.3. If F_i is a comparison $X=Y$ and Y is limited then X is limited.
4. A negated formula is unsafe **unless** it appears in a disjunct with one or more non-negated conjuncts and the free variables in the negated formula are limited as per rule 3. □

Fortunately, it turns out that these limitations do not compromise the expressiveness of our queries with respect to the algebraic relational operators with which users are typically aware and which we use as a reference (relational completeness).[21] Consequently further references to the DRC will refer to the safe-DRC unless explicitly noted otherwise. In particular, we will use the DRC and the value of an expression to formally define our concept of attribution.

4.2 Attribution and the DRC

We initially suggested that the intuition for attribution was somehow related to an interpretation for the logical expression of a query. We can now be slightly more specific about that idea. A relation r is the result of a query Q denoted by the DRC expression $\{X_1,X_2,\ldots,X_n \mid f(X_1,X_2,\ldots,X_n)\}$. The attribution of the tuples $(c_1,c_2,\ldots,c_n)$ of r when Q is evaluated on database d, denoted $Attr(r, (c_1,c_2,\ldots,c_n), Q, d)$ is related to the set of substitutions $f(c_1/X_1,c_2/X_2,\ldots,c_n/X_n)$ such that $I(f(c_1/X_1,c_2/X_2,\ldots,c_n/X_n)) = true$.

This must seem rather tautological. The attribution of a relation is somehow the relation itself. Therefore, we develop the idea by first considering attribution for *conjunctive queries* and then iteratively refining the model over progressively more general classes of queries.

4.3 Conjunctive queries

We begin the construction of our attribution model by first limiting the range of possible queries to the class of conjunctive queries (CQ). Were we to limit our model to conjunctive queries, attribution would still prove quite useful, for we know that CQ correspond to the class of all SQL queries constructed using selection-on-equality, project, and natural join (Maier 1983; Ullman 1988).

We define three different types of attribution for CQ expressions. After providing a definition for attribution equivalence, we confirm the equivalence of the attribution for equivalent CQ expressions. An algorithm for composing the attribution of an expression by iteratively drilling down through IDB is presented. We verify that composition produces the same attribution as the equivalent, unified query expressed only on EDB. Finally, we present some remarks on attribution granularity. We note the parallel between attributing some subset of values in a result and attributing using only some subset of values in the input sources.

[21] See (Ullman 1988 at 153) for a proof on the equivalence of the safe DRC and the relational algebra.

4.3.1 Attribution concept

We first define the term *conjunctive query* and then develop our model by considering our original intuition for attribution as a set of substitution lists for the variables in the expression.

Definition 4.7 Conjunctive query
A *conjunctive query* is an expression of the form:

$$\{X_1,X_2,\ldots,X_n \mid \exists Y_1,Y_2,\ldots,Y_m f(X\, X_1,X_2,\ldots,X_n,Y_1,Y_2,\ldots,Y_m)\}$$

constructed from a subset of the DRC, as defined earlier, consisting only of domain variables, constants, predicates that represent relations, conjunction, and existential quantification.[22] □

Example 4.7 Conjunctive queries
Conjunctive queries from the examples in Chapter 3 are:

E_1 = {HNAME | hotels(HNAME, ROOMS, PRICE)}
E_2 = {HNAME | hostels(HNAME, PRICE, STATION) $\wedge$ sites("Nakamise Dorsi", STATION)}
E_3 = {HNAME | regions(HNAME, REGION) $\wedge$ sites("Imperial Palace", REGION) $\wedge$ hotels(HNAME, ROOMS, PRICE)} □

If attribution consists of the set of substitutions for all variables in the expression, as demonstrated in Example 4.3, then attribution appears to combine a number of distinct concepts together at once. For example, there are a number of relations and variables used to determine the result that are not reflected in the set of free variables. Specifically, how do we know that the Imperial Palace and the hotels in our answer are in the same region? We need to know what region the Imperial Palace is in and what region each of the hotels are in. More generally, distinct information is conveyed in various subsets of the free and bound variables.

4.3.2 Types of attribution

Combinations of free and bound variables in the expression correspond to the intuition introduced in Chapter 3 that there are different types of attribution depending upon a particular user's interest. In this thesis, we will address three distinct subsets of the set of all variables and constants in an expression.

Perhaps the simplest attribution is that which we demonstrated in Section 4.3.1. From an intellectual property or remuneration perspective, knowing all of the values and variables used, irrespective of the role they play in answering the query, is significant.

[22] Because we can rewrite $r(X_1,X_2,\ldots c\ldots,X_n)$ as the formula $(r(X_1,X_2,\ldots Y\ldots,X_n) \wedge (Y = c))$ we see that conjunctive queries permit a safe or limited form of equality through multiple occurrences of the same variable in multiple conjuncts. See (Ullman 1988).

Definition 4.8 Comprehensive attribution
The *comprehensive attribution* for the relation represented by a CQ expression $r = \{X_1,X_2,\ldots,X_n \mid (\exists Y_1,Y_2,\ldots,Y_m)\ f(X_1,X_2,\ldots,X_n,Y_1,Y_2,\ldots,Y_m)\}$ is a set of pairs where each pair is a substitution list a for all of the variables in f that make f true, and the formula itself. We will sometimes write this as $\{f(a)\}$ or where p_i is a predicate in f and c_i is a constant, we might write $\{<p_i(c_i)>\}$ □

Note that for CQ expressions, a minimal list of substitutions must interpret every predicate in the expression as *true*. For an expression with $m + n$ variables, the substitution list must have $m + n$ substitutions.

For E1 in Example 4.7, a substitution a in the *comprehensive attribution* will provide values for the variables, HNAME, ROOMS, and PRICE. In addition to identifying all sources consulted in the query, both a unique substitution list and the set of lists convey additional information. In distinct substitution lists for CQ expressions, the same variable can recur in multiple predicates of the same formula. Multiple predicates correspond to multiple sources as in the case of an attribute used in a natural join. Note also that two distinct substitution lists might have the same values for all free variables X_i and differ in at most one existentially quantified variable Y_j hinting at the issue of multiple derivations raised in Chapter 3. From Example 4.7 we see that each answer in the result is attributable to two distinct substitutions. We will say more about multiple sources below.

A second type of attribution focuses on only the free variables in an expression rather than the set of all variables. Every occurrence of a free variable X_i in a distinct predicate p of a CQ corresponds to a *source* for X_i.

Definition 4.9 Source attribution
The *source attribution* for the relation represented by a CQ expression $r = \{X_1,X_2,\ldots,X_n \mid (\exists Y_1,Y_2,\ldots,Y_m)\ f(X_1,X_2,\ldots,X_n,Y_1,Y_2,\ldots,Y_m)\}$ is the set of pairs where each pair is a substitution list a for all variables in predicates of f that contain free variables and make f true, and the formula itself. □

A user interested in data quality characteristics of the answer that depend upon the sources from which the values in the answer are drawn, such as timeliness or accuracy, will examine the *source attribution* for the query result.

A third type of attribution concerns *relevant* sources. The quality of an answer to a query might depend not only upon values reflected in the result but also upon values used in evaluating query (restriction) conditions. We referred to this distinction in Chapter 3 as the difference between the quality of the answer to the query and the quality of a value in the answer.

The general intuition behind *relevant* substitutions is that omitting or changing one of these substitutions could increase or decrease the subset of domain values for any given free variable, corresponding to an attribute in the result.[23] In Example 4.7, were we to alter the condition "SNAME = 'Imperial Palace'" the query result would certainly differ.

Definition 4.10 Relevant attribution

The *relevant attribution* for the relation represented by a CQ expression $r = \{X_1,X_2,\ldots,X_n \mid (\exists Y_1,Y_2,\ldots,Y_m)\ f(X_1,X_2,\ldots,X_n,Y_1,Y_2,\ldots,Y_m)\}$ is the set of pairs where each pair is the formula f and a substitution list a for all *relevant* variables in f that make f true. We use the term *relevant* to capture constraints on the attribute domains represented by the variables in the head of the expression. All variables in the head (free in the formula for the expression) are relevant. In addition, a bound variable is relevant to the result if renaming the variable to some name not already in the expression (or eliminating a constant) would relax a constraint on one or more of the attribute domains in the result relation (free in the formula for the expression). □

Example 4.8

Consider again the CQ expressions from Example 4.7

E_1 = {HNAME | hotels(HNAME, ROOMS, PRICE)}
E_2 = {HNAME | hostels(HNAME, PRICE, STATION) $\wedge$ sites("Nakamise Dorsi", STATION)}
E_3 = {HNAME | regions(HNAME, REGION) $\wedge$ sites("Imperial Palace", REGION) $\wedge$ hotels(HNAME, ROOMS, PRICE)}

In addition, consider the more general CQ expression with domain variables as follows.

E_4 = {A | p(A,B,C) $\wedge$ q(C,D,D) $\wedge$ r(F,G,H) $\wedge$ s(H,J,J)}

Only HNAME is relevant in E_1. HNAME, a free variable, is relevant in E_2. STATION is also relevant in E_2. If we renamed the instance of STATION in relation sites, our new expression might appear as E_2 = {HNAME | hostels(HNAME, PRICE, STATION) $\wedge$ sites("Nakamise Dorsi", STATION2)} and the join condition would no longer constrain the possible values of HNAME. Likewise, the substitution "Nakamise Dorsi"/SNAME constrains values of HNAME by placing a bound on the values for STATION which in turn restrict HNAME. Only ROOMS and PRICE are not relevant in E_3. The quality of each answer in E_3 depends upon our knowledge of where the Imperial Palace is located in relation to each of the different hotels. In E_4, neither domain variable J nor domain variable H are relevant. Renaming one instance of H would alter the join condition between the relational predicates r and s. However, while these predicates pose an existence constraint on a tuple of the result set, they do not constrain the domain (and by extension, the quality) of values in the result set. □

We have defined *attribution* to provide variants on the different sources used to evaluate the answer to a query. However, there is often more than one way to ask a question. Likewise,

[23] Note explicitly the distinction between restriction conditions and existence conditions represented by Cartesian product. We say more about this in the discussion on *result granularity* below.

there are often different ways of answering the same question. In the next subsection, we consider *multiple derivations.*

4.3.3 Multiple derivations – the concept

The concept of multiple derivations addresses the observation that we can arrive at the same answer for the same question in different ways. First, we might ask the same question in different ways (equivalent queries). Second, a single query can produce identical answers in different ways.

Assuming the standard, containment-based definition for equivalent queries (Ullman 1989), we further divide the expression of equivalent query expressions into two categories: queries defined on a database comprised of base tables (Extension Data Bases EDB) and queries that also make use of relations defined in terms of other relations (Intensional Data Base IDB) (Ullman 1988). We refer to these as strict equivalence and composition respectively.

We call equivalent expressions defined on the same, extensional database strict equivalents. We first saw an example of strict equivalence in Example 3.5 of Chapter 3. Now, we adopt a more abstract representation to help generalize the concept.

Example 4.9 Strict Equivalence
Consider the following CQ expressions:

$E_5 = \{X \mid p(X,Y,Z) \wedge p(U,V,W)\}$
$E_6 = \{Z \mid p(Z,Y,U)\}$

Expressions E_5 and E_6 are syntactically different, yet they are equivalent. □

For equivalent queries defined using both intentional and extensional relations, we use the term composition. We first saw an example of Composition in Example 3.9 of Chapter 3. Now, we provide a more abstract representation.

Example 4.10 Composition
Consider the following CQ expressions:
Assume relation $s(U,V,W) \stackrel{def}{=} \{U,V,W \mid p(U,V,X) \wedge q(W,X,Y)\}$
and assume relation $E_7 = \{S,V \mid s(U,V,W) \wedge r(S,T,U)\}$.
We can then find a unifier such that $E_7' = \{S,V \mid p(U,V,X) \wedge q(W,X,Y) \wedge r(S,T,U)\}$. □

While equivalent queries lead to identical results, we might also think of a single expression producing identical results. For conjunctive queries, consider the same value appearing in different predicates as in the case of natural join. Natural joins in conjunctive queries were introduced in Q6 of Example 3.12 in Chapter 3. Here, we again offer a more abstract representation.

Example 4.11 Natural joins

$E_8 = \{X \mid p(V,W,X) \wedge q(X,Y,Z)\}$

Any substitution for the formula in E_8 must include *c/X*, suggesting that the relations represented by predicates *p* and *q* are both sources for *c*. □

Within a single relation we might also think of different sources when we consider non-key values recurring in multiple tuples. We spoke of weak equivalence and defined weak duplicates in Example 3.11 of Chapter 3. More generally, consider:

Example 4.12 Multiple instances of the same value
$E_9 = \{Y \mid p(X,Y,Z)\}$ where we assume no functional dependencies (or only the trivial dependency where XYZ → XYZ). Then, there may well be multiple values of Y corresponding to multiple tuples <X,Y,Z> in predicate p. □

In describing the issue of multiple derivations for an answer from the same query expression, we allude to the idea that a variable substitution might apply in more than one predicate, and that a single predicate may have duplicate, non-key values. Both issues suggest that there may be more to granularity than a list of substitutions in the formula for a query expression. We may be interested in a specific value (join-attribute or non-key value), and we may wish to distinguish between different substitutions in same attribution set. Issues of result and source granularity are addressed beginning in Section 4.3.7.

4.3.4 Multiple derivations from different expressions – strict equivalence

Our general intuition for attribution equivalence is that the substitutions are the same. In other words, equivalent comprehensive attributions should provide the same interpretation for the same expressions. Source and relevant attributions should be equally comparable. In the case of strict equivalence, if two conjunctive queries E_1 and E_2 are equivalent, then there is a containment mapping from E_1 to E_2 and from E_2 to E_1 (Ullman 1989). These containment mappings map predicates and variables between E_1 and E_2 and satisfy our intuitions about equivalent comprehensive, source, and relevant attributions. We therefore conclude that under different types of attribution, the attribution of equivalent CQ-expressions are equivalent.

First, we provide more formal definitions for what is meant by [comprehensive | source | relevant] attribution equivalence.

Definition 4.11 Attribution equivalence
Two attributions A_1 and A_2 are equivalent when there is a mapping for every variable and its corresponding predicates from A_1 to A_2 and from A_2 to A_1. □

Example 4.13 Attribution equivalence
Consider again Example 4.9.

$E_5 = \{X \mid p(X,Y,Z) \wedge p(U,V,W)\}$
$E_6 = \{Z \mid p(Z,Y,U)\}$

FORMAL MODEL

We say that the comprehensive attribution $A_{C5} \equiv A_{C6}$ because the containment mapping from E_5 to E_6 and vice versa, establishing the equivalence of E_5 and E_6, also maps the attribution substitutions.

The mapping establishing the equivalence of source attribution $A_{S5} \equiv A_{S6}$ is just the containment mapping for the free variables in E_5 and E_6. Likewise for the equivalence of relevant attribution $A_{R5} \equiv A_{R6}$. The mapping indicates that there is no free variable for a relational predicate p in E_5, that is not mapped in E_6. This will cause a problem once we add the union operator ($\cup$) into the query language. □

Given our definitions of equivalence, we then propose

Theorem 4.1 Attribution equivalence
If E_1 and E_2 are equivalent CQ expressions, then their [comprehensive | source] attributions, A_1 and A_2, are equivalent. If E_1 and E_2 are minimal, then attribution equivalence holds trivially for comprehensive, source, and relevant attribution.

Lemma 4.1 Comprehensive attributions of equivalent CQ expressions are equivalent.
This is trivially true by the definition of equivalence between E_1 and E_2.

Lemma 4.2 Source attributions of equivalent CQ expressions are equivalent.
Because the queries are equivalent, we know that the two expressions define the same relation. Therefore, in a CQ expression, the mapping must take the predicates containing free variables in E_1 to the predicates containing free variables in E_2 and vice versa.

Lemma 4.3 Relevant attributions of minimal, equivalent CQ expressions are equivalent.
We know that a mapping h from relevant variables in E_1 to variables in E_2 exists by equivalence. We need to verify that h maps all relevant variables in E_1 to relevant variables in E_2 and vice versa. From our definition of relevance, we know that we can exclude any redundant relational predicate as inherently irrelevant. Moreover, we know, from the query optimization literature, that removing redundant predicates from equivalent CQ expressions results in a unique, minimal equivalent CQ expression (Ullman 1989). As a consequence, the relevant attribution of equivalent CQ expressions is trivially equivalent because they are the same. Note that this claim assumes the absence of functional dependencies in the relation. If, for example, a relation has two disjoint candidate keys, then an expression that constrains one candidate key could be equivalent to an expression that constraints the second candidate key.

By Lemmas 4.1, 4.2, and 4.3, we conclude that the comprehensive and source attributions of equivalent queries is equivalent while the attributions of the minimal equivalents of equivalent queries are identical. □

4.3.5 Multiple derivations from different expressions, composition

A second way in which we get different expressions for the same query is when some predicates are defined in terms of others. As seen in Section 4.3.3, when we allow intentional databases (IDB), equivalent CQ-expressions can introduce new predicates and variables. We define the attribution of an expression involving IDB by rewriting the expression in terms only of the base data sources following the process of Unification in datalog queries (Ullman 1988).

The principle of composition establishes that, instead of re-writing the query, we may determine the attribution for composed queries in a recursive manner. First determine the attribution A in terms of both EDB and IDB. Extend each substitution $a_i \in A$ as follows. Treat every reference to an IDB as an independent CQ expression; extend a_i by attributing each IDB. For successive unfoldings, assuming that no recursive definitions are allowed, we eventually arrive at the attribution for the initial expression in terms of base data sources.

Example 4.14 Attribution composition

$$E_1 = p \wedge q \wedge r$$
$$r \overset{\text{def}}{=} E_2 = s \wedge t$$
$$E_3 = p \wedge q \wedge s \wedge t$$

Step 1. Get the attribution for E_1 in terms of p, q, and r.
Step 2. Project the substitution list from Step 1. onto r.
Step 3. Attribute Step 2. on the expression for r.
Step 4. Combine the attribution from Step 3. to the attribution from Step 1. □

More generally, we propose a CQ expression E' for a database d of the form:

$$p_1 \wedge p_2 \wedge \ldots \wedge p_n \wedge q_1 \wedge q_2 \wedge \ldots \wedge q_m$$

where p_i is a predicate for a relation $r_i \in d$ and $\forall j$, q_j is a predicate for a relation $r_j \notin d$ and q_j is defined by a CQ-expression over predicates p_i.

Definition 4.12 Attribution of a composed expression

The attribution of the result r from E' defined on d' in terms only of relations in d, is defined as $attr(r, E, d)$ where d explicitly excludes $\forall j$, predicates q_j and E is the re-write of E' in terms of d. □

It follows that we can build progressively deeper layers of indirection by defining a set of predicates r_k defined in terms of p_i's and q_j's and so forth resulting in correspondingly more complex re-writes.

While re-writing provides us with a consistent definition for the attribution of expressions in the presence of views and base relations, it presents some pragmatic challenges. Neither user nor system may initially be aware of underlying data sources. Users may be uninterested in pursuing the attribution of certain intermediate-level sources. Rather than re-writing the entire query a priori, we would prefer to attribute by iteratively unfolding successive layers of IDB as necessary.

Algorithm 4.1 Attribution composition

```
Compose $(A, f)$ {                                                   (1)
if $f$ has no $q$'s then return $A$                                   (2)
else pick $q_i$, an IDB in $f$                                        (3)
    $f := p_1 \wedge p_2 \wedge \ldots \wedge q_{i-1} \wedge q_{i+1} \ldots q_m$          (4)
    Compose (Unfold $(A, q_i), f)$ }                                  (5)

Unfold $(A, q)$ {                                                     (6)
if $A$ is $\varnothing$ then return { }                               (7)
else pick $(a,f) \in A$                                               (8)
    let $g$ be the formula for IDB $E$ representing $q$               (9)
    let $u$ be the unifier for $h = unify(f,g)$                       (10)
    let $E'$ be $E$ as defined by $g$ with the renaming of $u$        (11)
    $B =$ attr( $a(q)/x, E', d'$)                                     (12)
    Rewrite $(B, u(a - a(q)), h) \cup$ Unfold $(A - \{(a,f)\}, q)$ }  (13)

Rewrite $(B, a, h)$ {                                                 (14)
if $B$ is $\varnothing$ then return { }                               (15)
else pick $(b,g) \in B$                                               (16)
    $\{<\{a \circ b\},h>\} \cup$ Rewrite $(B - \{(b,g)\}, a, h)$     (17)     □
```

We use Compose (A, f) to recurse through the IDB in f. For each IDB, we find the definition for the IDB in line (9) and find a unifier in line (10) to be certain that we can rename variables appropriately. In line (11), we rewrite the expression for the IDB accounting for the variable renaming and call this E'. Finally, we attribute the specific tuple in the IDB by pushing constants from the original substitution into the corresponding variables of E'. We denote this as $E'(a(q)/x)$ in line (12). Because this attribution itself returns a set of substitution - formula pairs, we replace the original substitution in A with the set of substitutions from the attribution of E'. Note in line (17) where the set of new pairs uses the unified formula h and combines the original substitution a with substitutions for the IDB b. Line (13) simply removes the duplicate substitutions.

Theorem 4.2 Attribution composition
Attribution composition computes the attribution of a composed expression.

Assume without loss of generality the following CQ expressions E_1, E_2, E_3 defined by the formulas f, g, and h respectively s.t.

$E_1 \overset{\text{def}}{=} f = (p_1 \wedge p_2 \wedge \ldots \wedge p_n \wedge q)$ where q is the only IDB in E_1
$E_2 = q \overset{\text{def}}{=} g = (r_1 \wedge r_2 \wedge \ldots \wedge r_m)$ where $r_i \in d$
$E_3 \overset{\text{def}}{=} h = (p_1 \wedge p_2 \wedge \ldots \wedge p_n \wedge r_1 \wedge r_2 \wedge \ldots \wedge r_m)$

Note that the formula g for E_2 already has variables renamed and reordered as in Line (10) and (11) so that references to E_2 in the proof below correspond to E' in Line (12).

Given E_1 defined on $d' = d \cup \{q\}$ and r, the result of evaluating E_1 on d', attribution composition computes the [comprehensive | source | relevant] attribution of result r in terms of d as defined by attr(r, E_3, d).

Lemma 4.4 **$(a_3,h) \in A_3$ is a comprehensive attribution for E_3 if and only if $(a_3,h) \in$ Compose(E_1,f) .**

Case ($\rightarrow$)
Pick a random substitution $(a_3,h) \in A_3$ and split it: Project a_3 onto f and g.
We know that $a_3(f) = a_1 \in A_1$ because $\forall i$, $I(p_i(a_3(p_i)/x)) = true$ and $\forall j$, $I(r_j(a_3(r_j)/x)) = true$ where q is defined by the r_j's. Similarly, we know that $a_3(g) = a_2 \in A_2' \subseteq A_2$ where A_2' as the attribution for the tuple defined by $a_1 \cap a_2$, a tuple in q.
Compose passes A_1 to **Unfold**. **Unfold** calls attr($\{a_3(q)\}$, E_2, d') which looks for substitutions of E_2 with $a_3(q)$ pushed into the expression. Attr($\{a_3(q)\}$, E_2, d') is $A_2' \subseteq A_2$ because the attribution of $E_2 = A_2$ and $a_1(g)$ makes E_2 true therefore $A_2' \subseteq A_2$.
Unfold is applied to every value of A_1 so certainly it calls itself on a_1.
Unfold calls **Rewrite** with a_1 and A_2'.
Rewrite is applied to every element of A_2' so certainly is is applied to a_2.
But **Rewrite** takes h, the unification of f and g, and returns $a_1 \circ a_2$ which is a_3.
Case ($\leftarrow$)
If (unify(f,g)) = h, does every pair $(a_1 \circ a_2, h)$ appear as a substitution in A_3? Pick some arbitrary a_1 from a pair in A_1. Now we cannot pick just any a_2. **Compose** creates A_2' from attr($\{a_1(g)\}$, E_2, d'). So pick any a_2 from a pair $\in A_2'$. We know $a_1 \circ a_2$ paired with h appears in A_3 if it makes E_3 true. $\forall i$, $I(p_i(a_1(p_i)/x)) = true$ and $\forall j$, $I(r_j(a_2(r_j)/x)) = true$. But are E_1 and E_2 true at the same time (i.e. do they make h true)? Because we know a_2 is from a pair in A_2' by construction, we know that a_2 makes E_2 true for a true interpretation of A_1. Therefore, we know that $(a_1 \circ a_2, h) \in A_3$. □

Lemma 4.5 **$(a_3,h) \in A_3$ is a source attribution for E_3 if and only if $(a_3,h) \in$ Compose (A_1,f) where A_1 is the source attribution for E_1.**

Case ($\rightarrow$)
Pick a random source attribution $(a_3,h) \in A_3$ and split it: Project a_3 onto f and g. These are just the free variables in E_3 and accompanying variables that identify unique instances of

tuples containing a particular value for a free variable. We know as before that $a_3(f) = (a_1,f) \in A_1$ because for the predicates p_i, $E_1 \subseteq E_3$ and for predicate q, because q is defined in terms of the predicates r_j, we know that the free variables for q are also assigned in $a_3(f)$. Similarly, we know again that $a_3(g) = (a_2,g) \in A_2' \subseteq A_2$ where A_2' as the attribution for the tuple defined by $a_1 \cap a_2$, a tuple in q.
Compose passes A_1 to **Unfold** with $q \overset{\text{def}}{=} E_2$ with formula g.
Unfold is called on every value of A_1 so certainly it is called on a_1.
Unfold calls attr($\{a_1(g)\}$, E_2, d') which we know is $A_2' \subseteq A_2$ because the attribution of $E_2 = A_2$ and $a_1(g)$ makes E_2 true therefore $A_2' \subseteq A_2$.
Rewrite is called for every element of A_2' so certainly it is called for a_2.
But **Rewrite** takes h, the unification of f and g and returns $a_1 \circ a_2$ which is a_3.

Case (←)
If (unify(f,g)) = h, Does every pair ($a_1 \circ a_2$, h) appear as a substitution to a relational predicate containing a free variable in A_3? Pick some arbitrary a_1 from a pair in A_1. Now we cannot pick just any a_2. **Compose** creates A_2' = attr($\{a_1(g)\}$, E_2, d'). So pick any a_2 from a pair $\in A$-$_2'$. We know $a_1 \circ a_2$ paired with h appears in A_3 if it makes E_3 true. $\forall i$, $I(p_i(a_1(p_i)/x)) = true$ and $\forall j$, $I(r_j(a_2(r_j)/x)) = true$. But are E_1 and E_2 true at the same time (i.e. do they make h true)? Because we know a_2 is from a pair in A_2' by construction, we know that a_2 makes E_2 true for a true interpretation of A_1. Therefore, we know that $(a_1 \circ a_2, h) \in A_3$. □

Lemma 4.6 $(a_3,h) \in A_3$ is a relevant attribution for E_3 if and only if $(a_3,h) \in$ Compose (A_1,f). Where A_1 is the relevant attribution for E_1.

This is more complicated because we need to verify that $relevant(E_3) = relevant(E_1) + relevant(E_2)'$ where $relevant(E_2)' \subseteq relevant(E_2)$ and $relevant(E)$ refers to the relevant variables in E and likewise for $free(E)$; $bound(E)$. We form $relevant(E_2)'$ as we formed A_2' previously. We attribute only the relevant variables in q on the expression E_2. For convenience, we assume that the CQ expression is minimal.

Case (→)
Pick some relevant attribution $(a_3,h) \in A_3$ and split it: Project a_3 onto f and g.
We need to establish that $a_3(f) = (a_1,f) \in A_1$ and $a_3(g) = (a_2,g) \in A_2' \subseteq A_2$ where A_2' as the attribution for the tuple defined by $a_1 \cap a_2$, a tuple in q.
A substitution c/X is in a substitution list for a_3 because either X is free in E_3 or c/X joins two relational predicates, at least one of which is recursively joined to a relational predicate containing a free variable or is a constant from the original query expression that appears in a relational predicate recursively joined to a predicate containing a free variable of E_3.

Case 1. $X \in relevant(E_3)$ and $X \in free(E_3)$. $Free(E_3) = X \in free(E_1) \subseteq relevant(E_1)$ by definition of the equivalence of E_1 and E_3.

for $X \in free(E_3) = Y \in free(E_2)$, Y must also be free in E_1 because E_2 is q in E_1 (e.g. $Y \in free(E_2) \rightarrow Y \in free(E_1)$). Consequently, at least for the relevant variables in E_2 that are free, we know $a_2 \in A_2' \subseteq A_2$

Case 2. $X \in relevant(E_3)$ joins relational predicates to a recursively joined set of relational predicates or X constrains one predicate in a recursively joined set of relational predicates (e.g. X is a constant or X appears multiple times in a single relation). All such predicates are in the set p_i and at least one joined predicate contains a free variable in h. Then X is relevant in E_1 so $a_3(f) = a_1$ for $(a_1,f) \in A_1$.

Case 3. $X \in relevant(E_3)$ is like Case 2 except all such predicates are in the set r_j. Then X is relevant in E_2 so $a_3(g) = (a_2,g) \in A_2' \subseteq A_2$ where A_2' as the attribution for the tuple defined by $a_1 \cap a_2$, a tuple in q. (Recall that $g \overset{\text{def}}{=} q$ in E_1).

Case 4. $X \in relevant(E_3)$ appears in both some predicate p_i and some predicate r_j. Then X must appear in predicate q of E_1 (X is not free in E_1 so it must be bound in E_1 and appear in q in order to appear in both p_i and r_j in E_3). Therefore X is relevant in E_1 so $a_3(f) = a_1$ for $(a_1,f) \in A_1$.

It is possible that $a_3(g)$ is empty, which occurs when we consider a Cartesian Product and then do not restrict variables between arguments to the Cartesian Product. In this case, attribution relevance is trivially true in the p_i's.

From here, we do the same unfolding as before and conclude that **Compose** returns $a_1 \circ a_2$ which is a_3.

Case (←)
Pick some random substitution list $a_1 \circ a_2$ as before and verify that $(a_1 \circ a_2,h) \in A_3$.
Proof by contradiction. Suppose not. Then there must be a substitution $c/X \in a_1 \circ a_2$, $c/X \notin a_3$, or $c/Y \in a_3$, $c/Y \notin a_1 \circ a_2$.

Case 1. Pick Y. If Y is relevant in E_3, then Y must constrain the free variables in h in some way.
If Y is a free variable, then $c/Y \in a_1$, a contradiction.
If Y constrains a predicate containing a free variable through some recursively joined set of predicates amongst the p_i's, then $c/Y \in a_1$, a contradiction.
If Y constrains a predicate containing a free variable through some recursively joined set of predicates amongst the r_j's and is $a_1 \circ a_2$ as assumed above, then $c/Y \in a_2$, a contradiction.
If Y is relevant and appears in both some predicate p and some predicate r then $c/Y \in a_1$, a contradiction (see Case 4 for the (→) direction).
Therefore, by contradiction, we conclude that there is no $c/Y \in a_3$, $c/Y \notin a_1 \circ a_2$.

Case 2. Pick X.
If $c/X \in a_1$ because it is a free variable in E_1, then by definition, $c/X \in a_3$, a contradiction.
If $c/X \in a_1$ because it constrains a free variable through predicates p_i, then $c/X \in a_3$, a contradiction.
If $c/X \in a_1$ and appears in both in q's and p's, then $c/X \in a_3$, a contradiction.
Now we need to be careful. Remember that a_2 is selected from attributing $a_1(g)$. It is possible for $a_1(g) = \{\}$ as in the case of Cartesian Product. attr($\{a_1(g)\}$, E_2, d') is non-empty only when there is a relevant variable in q.
If $c/X \in a_2$ because it is free in E_2 and free in E_1, then we know $c/X \in a_3$, a contradiction.
If $c/X \in a_2$ because it is free in E_2 and bound and relevant in E_1, then we know $c/X \in a_3$, a contradiction.
If $c/X \in a_2$ because it is bound in E_2, occurs among the predicates r_j and constraints a free variable in E_2 that is relevant in E_1 (through predicate q in E_1), then we know $c/X \in a_3$, a contradiction. □

Therefore, we conclude that attribution composition computes the attribution of a composed expression. □

It is important to note the subtlety required in composing relevant attribution. Our definition of relevance depends upon drawing a distinction between constraints on attribute domains and explicit query syntax. We saw some challenges for managing relevant attribution in Example 3.9 of Chapter 3. Consider, more generally, two equivalent queries where selections are pushed down in one case but not in the other.

Example 4.15 Composing relevant attribution

$E_{10} = \{A \mid p(ABCDEF) \wedge s(FGH)\}$
$E_{11} = \{A \mid q(ABC) \wedge r(DEF) \wedge s(FGH)\}$
where $p(ABCDEF) = q(ABC) \wedge r(DEF)$

Syntactically, we observe that F is relevant to A in E_{10} but not in E_{11}. Yet, the equivalence of E_{10} and E_{11} confirms that F indeed does not constrain values of A in the result.[24] □

Theorem 4.2 confirms our intuitions about how attribution should work in the context of composed queries. It indicates that, at least for conjunctive queries, we can recursively drill down through progressive layers of indirection. More generally, Theorem 4.1 and Theorem 4.2 together allow us to conclude that, though there are many different ways to construct a CQ expression, comprehensive, source, and relevant attributions for equivalent CQ expressions are equivalent.

[24] Note that we are essentially saying that composition holds for relevant attribution because we explicitly define composition and relevance that way.

4.3.6 Multiple derivations within a single expression

We saw in Section 4.3.3 how different substitutions might correspond to the same values within a single expression. Both multiple occurrences of a single variable and multiple substitutions proving the same result are modeled in a straightforward manner.

Multiple occurrences of a variable between expressions, as in the concept of relevant variables, are consistent with the semantics of algebraic natural join. That a single variable appears as a join attribute suggests that it derives from two or more distinct relations in a single expression. See Example 4.11. We will say more about what it means to derive from a relation rather than from a substitution in our discussion of granularity to follow.

In addition to identifying duplicate values through multiple occurrences of a single variable in an expression, non-key values can repeat in different facts of a single predicate corresponding to different tuples of a single relation as in Example 4.12.

Rather than being problematic, however, we believe that this highlights a benefit of using substitutions to define attribution. Duplicate values suggest an opportunity for users to explicitly identify either a specific instance of a value or all such instances. In the relational data model we know that we can identify specific instances through functional dependencies. That our attribution model draws a distinction between specific instances of a value and all such instances introduces the concept of granularity.

4.3.7 Granularity – the concept

The intuition behind granularity is that attribution is simply a pointer from query results to query sources. Granularity addresses the precision with which the pointer identifies data in a source or in a result. Source granularity allows the user to receive a list of references that provides greater (or less) detail. Note that a substitution, defined as a list of value-substitutions and the formula to which the substitutions are applied, implicitly associates values with one or more relations. As a consequence, rather than a substitution value, we might return the tuple(s) containing a value or even the relation name. Source granularity was first discussed in Examples 3.15 and 3.16 of Chapter 3. More abstractly, consider:

Example 4.16 Source granularity

$E_{12} = \{A, E, F \mid p(A, B, C) \wedge q(C, D, E) \wedge r(F, G, H)\}$

where the source of interest is represented by predicate $p(A, B, C)$
if a is a substitution list for the formula of E_{12} then $a(p) = \langle c_1/A, c_2/B, c_3/C \rangle$ and the substitutions make predicate p true. We can think of a specific tuple instance as a source for the evaluation of E_{12}, (c_1, c_2, c_3). At the opposite extreme, we might roll-up all such tuple references by identifying the relation for predicate p as a source. The two poles define a continuum where, using the notation loosely, we can specify some tighter bound on tuples from the base relation that are used to evaluate the result. Consider, for example, $(c_1, \_, \_)$ as the set of all tuples in the relation for predicate a subset of tuples in the relation for predicate p where the value of the first attribute is c_1. □

Similarly, result granularity allows the user to ask attribution questions to varying degrees of specificity. Initially, we assumed that attribution applied to a query result as a whole. Implicitly, however, we accepted the notion that users might have an interest in only one portion of the result. Indeed our algorithm for attribution composition exploits the fact that we can attribute parts of relations. Rather than asking for the attribution of a relation defined by an expression, we may wish to know the attribution for a specific tuple, column, or value. Example 3.17 of Chapter 3 offered a first example of result granularity.

Example 4.17 Result granularity
Consider

$E_{13} = \{A, B, E \mid p(A, B, C) \wedge q(C, D, E)\}$

Again using the notation loosely, we might demonstrate an interest only in tuples where the value for variable B is c_2 (denoted $(\_, c_2/B, \_)$). For example, all students in a student database who have the last name "Smith." At the extreme, we might wish to attribute only a single, specific tuple $(c_1/A, c_2/B, c_3/C)$. □

We can therefore think of a query result as a relation and the attribution of that result as the corresponding input relations. However, being able to specify different granularities is useful because it enables precision while at the same time introducing possible efficiencies. When we attribute a relation, we do not necessarily know which substitutions correspond to specific values in the relation. Intuitively, every value is the result of distinct substitutions. If such exactitude is not necessary, however, as in the case of the list of references at the end of a text, attributing a group of values to a single list of relation names reduces the amount of necessary attribution metadata.

4.3.8 Source granularity

In source granularity, we vary the precision with which we identify the formula and the one or more corresponding variable substitutions that together define an attribution substitution. We hinted at source granularity when we discussed multiple derivations within a single expression. In particular, a single substitution may occur in multiple predicates. Multiple facts (with the same non-key attribute values) may correspond to a single value substitution.

Our definitions for different types of attribution correspond implicitly to different source granularities. Comprehensive attribution gives the complete list of substitutions for defining one true interpretation of a CQ expression. Source attribution identifies explicit tuples but only in relations from which free variables are drawn. Relevant attribution defines sets of tuples for selected predicates in the expression.

Example 4.18 Source granules and attribution substitutions
Consider again DRC 2.1 from Chapter 3.

DRC2.1 {HNAME | regions(HNAME, REGION) ∧ sites("Imperial Palace", REGION) ∧ hotels(HNAME, ROOMS, PRICE)}

The comprehensive attribution for the expression is the set of pairs:

{<*f*("Imperial"/HNAME, "Hibiya"/REGION, "Imperial Palace"/SNAME, "single"/ROOMS, 34000/PRICE)>;
<*f*("Imperial"/HNAME, "Hibiya"/REGION, "Imperial Palace"/SNAME, "double"/ROOMS, 39000/PRICE) >;
<*f*("Dai-Ichi"/HNAME, "Hibiya"/REGION, "Imperial Palace"/SNAME, "single"/ROOMS, 10000/PRICE) >;
<*f*("Dai-Ichi"/HNAME, "Hibiya"/REGION, "Imperial Palace"/SNAME, "double"/ROOMS, 80000/PRICE) >}

As illustrated above, projecting a substitution list onto a relational predicate in *f* returns a tuple that appears in the corresponding relation.

By contrast, consider the relevant attribution:

{<*f*("Imperial"/HNAME, "Hibiya"/REGION, "Imperial Palace"/SNAME)>;
<*f*("Dai-Ichi"/HNAME, "Hibiya"/REGION, "Imperial Palace"/SNAME) >}

Projecting one of the substitution lists onto the predicate hotels returns only the substitution "Imperial"/HNAME which we can apply as:

hotels("Imperial"/HNAME, PRICE, ROOMS) and corresponds to two tuples:
("Imperial", "single", 34000) and ("Imperial, "double", 39000) □

Example 4.18 suggests the ambiguity that can occur in attribution where multiple instances of a value in a source may contribute to a single answer. The ambiguity also offers flexibility, however. Individual variable substitutions indicate all occurrences of one or more variables in an expression whereas attributing with source tuples directs the attribution to identify explicit instances. Note that our use of tuple-level source granularity is a proxy for identifying unique instances. Leveraging functional dependencies may provide additional value here. Buneman et al. also hints at the potential of using functional dependencies in attribution and addresses the issue of unique instances for their more general deterministic semistructured data model (Buneman 01).

We note that an arbitrary granule defines a subset of values in a source (or result) relation. Specifying an arbitrary source granule does not imply that all valid substitutions for the expression are contained within the granule. Likewise, not every substitution within a coarse granule of a CQ expression may give a true interpretation for the expression.

Example 4.19 Interpreting source granules in attribution

Consider a variant on DRC 2.1 from Chapter 3 where we ask for "single" rooms by the "Imperial Palace."

DRC2.1' {HNAME | regions(HNAME, REGION) ∧ sites("Imperial Palace", REGION) ∧ hotels(HNAME, "single", PRICE)}

The comprehensive substitutions are now:

{<f("Imperial"/HNAME, "Hibiya"/REGION, "Imperial Palace"/SNAME, "single"/ROOMS, 34000/PRICE)>;
<f("Dai-Ichi"/HNAME, "Hibiya"/REGION, "Imperial Palace"/SNAME, "single"/ROOMS, 10000/PRICE) >}

The corresponding source attribution is:

{<f("Imperial"/HNAME, "Hibiya"/REGION, "single"/ROOMS, 34000/PRICE)>;
<f("Dai-Ichi"/HNAME, "Hibiya"/REGION, "single"/ROOMS, 10000/PRICE) >}

For coarse grained source attribution, we might identify a granule using only the source substitutions:

{<f("Imperial"/HNAME)>;
<f("Dai-Ichi"/HNAME) >}

Applied to the predicate hotels, we know that the substitution "Imperial"/HNAME corresponds to two tuples:
("Imperial", "single", 34000) and ("Imperial, "double", 39000)

However, the second of the two tuples does not produce a valid interpretation of the original query expression.

Definition 4.13 Source granularity
A source granule on a relation, denoted by predicate p in a CQ expression E, is defined by a CQ expression on predicate p (i.e., it is a view). □

Observation 4.1 Defining a source granule in terms of substitutions
Although an arbitrary source granule need not include all valid tuples for evaluating the truth of an expression, suppose that we have an expression E with attribution A. If we define our source granules using the substitution $a \in A$, we are assured that the source granules will always contain at least those tuples necessary to evaluate the query and produce the result corresponding to the attribution. □

Example 4.20 Defining a source granule in terms of substitutions
Suppose that we had the attribution A for a query expression E with formula f. f includes the relational predicate p such that for some substitution list $a \in A$, $a(p) = c_1, \ldots, c_n$ and c_i/X_i where X_i is a domain variable in p. We can then define a source granule for p as a query expression $\{Y_1, \ldots ,Y_m| \, p(Y_1,\ldots, Y_m)\}$ where we substitute c_i/Y_j as appropriate (e.g. where $X_i = Y_j$). The source granule therefore describes p', a tighter bound on p that still is guaranteed to contain at least those tuples that satisfy the original expression E. □

Tuple-level granularity constitutes a value/variable substitution for every argument in a relational predicate and describes a specific instance of a source relation. As noted above, although we define attribution in terms of substitutions, comprehensive and source attribution provide tuple-level granularity. Assuming no functional dependencies, assigning a value to each domain variable in a relation uniquely identifies an instance of the relation. Substitution-level granularity, such as is used in our definition of relevant attribution, implicitly includes every tuple from each constituent base relation that includes a particular attribute-value/domain variable substitution. At the extreme, we can speak of a relation-level source granule as simply a relation name. At the extreme, rather than attributing with specific substitutions, we can simply provide relation names as a proxy for all tuples in the corresponding relation.

In general, tuple-level substitutions are the finest grained (most specific), and relation-level granules are the most coarse, across all attribution types. This says that, where identifying specific values or instances of values is unimportant, we can always attribute with more general relations. For purposes of intellectual property or remuneration, for example, knowing the relation names may be sufficient. Likewise, for data quality purposes, knowing the relation may be enough to convey information about reputability. By contrast, verifying or correcting anomalous values may require finer granularity.

If we limit granules to those defined by substitutions, then we may make the following two observations about the relationship between different levels of source granularity

Observation 4.2 Generalizing from fine- to coarse-grained source granules
Given a set of source [comprehensive | source | relevant] substitutions that constitute a particular degree of specificity, we may always compose a query over the source granules that will contain at least the original substitutions. At the limit, we can always define a source granule that contains the original substitutions as the original base relation(s). □

Observation 4.3 Specializing from coarse- to fine-grained source granules
Assuming a set of [comprehensive | source | relevant] substitutions that constitute a particular degree of specificity, we may always re-attribute the same query expression and query result and return source granules that contain no more than the original set of substitutions. At the limit, we know that the tightest bound is the set of exactly those comprehensive, source, or relevant tuples that evaluate the expression to true. □

Because we define granularity as a composed query on a source predicate p, we may also make the following observations about the implications of varying source granularity on other properties of attribution.

Observation 4.4 Attribution composition is preserved
We define source granules in terms of composed queries on the base sources. Source granules therefore implicitly constitute IDB. At the extremes, either a source granule contains exactly those tuples that evaluate the expression to true or it is the identity on the EDB (i.e. relation-level source granularity). We already know that we can compose tuple-level substitutions. At the opposite extreme, if we attribute with a source relation name rather than

a set of source substitutions, we know that we can unfold by composing the relation names of the relations used to construct an IDB. □

Observation 4.5 Attribution of strictly equivalent queries is preserved
For relevant attribution, this is again, trivial. There is a unique minimal equivalent; regardless of the source granularity used, the relevant attribution is identical. For comprehensive and source attribution, we may again rely upon the containment map between equivalent expressions. Because the variables map to one another in the same predicates, we are assured that a source granule in one expression, defined as a query composed on a predicate, prescribes the same subset of base relation tuples in the equivalent expression. □

4.3.9 Result granularity

Result granularity stems from two observations. First, from the beginning, we intuited that users may have some interest in greater precision than simply attributing the result of a query. One tuple or even one value may raise particular interest. We refer explicitly to result granularity in our definition of composition. To compose an attribution recursively, we attribute substitutions in a predicate, not the entire relation represented by the predicate.

A second observation motivating result granularity stems from relational closure and the fact that relational query answers can serve as inputs to subsequent queries. As a consequence, source granularity issues like "all occurrences of a value" or "the specific instance of a value" may apply equally to results as well as to sources.

Example 4.21 Result granularity
Consider a variant on DRC 2.1 from Chapter 3.

```
DRC2.3    {HNAME, PRICE | regions(HNAME, REGION) ∧ sites("Imperial Palace", REGION) ∧
          hotels(HNAME, ROOMS, PRICE)}
```

We know that the result set includes the "Imperial, 34000" the "Dai-Ichi, 80000" and the "Dai-Ichi, 10000". A user might only have an interest in the "Dai-Ichi" hotel rather than the "Imperial". A different user might only be interested in the attribution for values of PRICE. □

The concept of result granules is consistent with our definitions of attribution, which refer to the substitutions that make the expression for the result true. As with source granules, we can imagine attributing the specific instance of a value in the result rather than all instances of a value. In Chapter 3 we saw how the projection of a non-key attribute from a base relation can result in multiple sources for the same value.

Mindful that an IDB is simply the result of a query[25], we follow our definition of source granules in defining result granules.

Definition 4.14 Result granularity
A granule of result *r*, defined by CQ expression *E* evaluated on database *d* is a result *r'* defined by a CQ expression *E'* composed on *E* for database *d*.[26] □

Observation 4.6 Attribution of a result tuple
It follows from our definition of a result granule that the attribution of a specific tuple *t* in a result *r*, assuming no knowledge of functional dependencies, is then simply the attribution of composing a query on the result *r* for the specific tuple of interest.

Moreover, because we define result granules using query composition, we are assured of

Observation 4.7 Attribution of strictly equivalent queries is preserved.
We already know that the comprehensive and source attribution of strictly equivalent queries is equivalent. Because this equivalence is preserved over composition, we conclude that the attribution of an arbitrary result granule is equivalent given equivalent CQ expressions. □

To define the relationship between result granules and source granules we offer the following observations.

Observation 4.8 Attribution of a result tuple
For relation-level source granules, the attribution of one tuple in the result of a CQ expression is the same as any other tuple in the same result. This merely conforms to the intuition that in a CQ expression, every conjunct applies equally to every tuple. □

Observation 4.9 Comprehensive attribution of result values
For comprehensive attribution, we may make the following stronger claims. First, regardless of source granularity, we observe that the comprehensive attribution for one value in a result tuple is the same as that for every other value in the same tuple. Second, if we limit ourselves relation-level source granules, the comprehensive attribution for a value in the result is the same as that for every other value in the result. □

The relationships between different granules has particular relevance for practical implementation, because it promises significant reductions in the amount of attribution metadata necessary to satisfy different user objectives.

[25] The closure property of relational theory dictates that a query result (output) may in turn serve as a source (input) to some other expression (Maier 1983; Ullman 1988)

[26] As we enrich our query language, we will eventually define a source or result granule by composing any positive query on a source or result relation, respectively.

4.4 Adding theta comparisons

We now move to refine our theory by extending the richness of the query language. The introduction of theta comparisons challenges some of our earlier conclusions about attribution when limited to CQ expressions. However, we verify that, for strictly equivalent queries, the comprehensive and source attribution of equivalent queries remains equivalent. Moreover, we conclude that for all types of attribution, attribution composition continues to hold.

4.4.1 Attribution concept

The first language extension introduces arithmetic comparisons in atoms of the form $(X\theta Y)$, $(X\theta c)$, or $(c\theta X)$ where c is a constant and X and Y are either free or bound variables that are limited in the manner defined for the DRC above. We refer to our extended queries as CQT expressions (or CQ expressions with theta comparisons). The set of θ operators are $\{<, \leq, \geq, >, \text{and} \neq\}$. For current purposes, we exclude explicit equality from the set of comparisons; explicit equality is incorporated into the language independently.[27]

Example 4.22 θ-comparison
First, consider a variant on query Q2 of Chapter 3.

E_{14} = {HNAME, PRICE | regions(HNAME, REGION) ∧ sites("Imperial Palace", REGION) ∧ hotels(HNAME, "single", PRICE) ∧ (PRICE < 15000)}

Second, we present a more abstract case.

E_{15} = {W | p(V,W,X) ∧ q(X,Y,Z) ∧ (V > 10)} □

Extending our definitions of attribution from CQ-expressions, we see that the introduction of comparisons does not introduce new relational predicates but may introduce new variables or perhaps constants for comparison. To better understand the implications of these changes for our theory, we revisit our analysis for conjunctive queries beginning with types of attribution.

4.4.2 Types of attribution

We initially defined comprehensive attribution as a set of substitution lists for all variables in the expression applied to the formula for the expression itself such that the interpretation of the formula is true. The definition for comprehensive attribution remains unchanged.

While θ-comparisons may introduce new variables into the expressions, under the limitations of safety, every variable is still *limited* in the sense that it must appear in a (non-negated) relational predicate. Consistent with Definition 4.9 on source attribution and Definition 4.13 on source granularity, non-predicate atoms are not considered sources. For E_{14} above, the arithmetic comparison is not considered a source for PRICE.

[27] Recall that conjunctive queries already included a "safe" or limited version of equality-comparisons. See note and text at 9.

Likewise, Definition 4.10 for relevant attribution remains unchanged. The introduction of comparisons, however, does provide new alternatives for constraining the domain of a free variable. In E_{15}, V is relevant to W.

4.4.3 Multiple derivations from different expressions, strict equivalence

The same two categories for multiple derivations that we identified in CQ expressions apply when theta-comparisons are added. Multiple derivations may stem from equivalent expressions or from multiple occurrences within a single expression. For equivalent expressions on the same database, we now need to consider containment not only between predicates of equivalent expressions but between non-predicate atoms as well.

Example 4.23 Multiple derivations

$E_{16} = \{XY \mid p(XYZ) \wedge q(UVW) \wedge (X \neq U) \wedge (X \leq U)\}$

$E_{17} = \{XY \mid p(XYZ) \wedge q(UVW) \wedge (X < U)\}$ □

The problem is tied to the introduction of new atoms in the form of theta comparison. The relationship between arithmetic comparisons of equivalent queries is not always clear as indicated in the following example from Ullman (1989).

Example 4.24 Interactions between arithmetic comparisons and relational predicates

$E_{18} = \{XY \mid p(XYZ) \wedge q(UV) \wedge (U \leq V)\}$

$E_{19} = \{XY \mid p(XYZ) \wedge q(UV) \wedge q(VU)\}$ □

Fortunately, we do know that a containment mapping does hold between the relational predicates in equivalent CQT expressions (Ullman 1989). Furthermore, the property of safety guarantees that all domain variables are captured in the containment mapping.

Theorem 4.3 Attribution equivalence
If E_1 and E_2 are equivalent CQT expressions, then their [comprehensive | source] attributions, A_1 and A_2, are equivalent.

Lemma 4.7 Comprehensive attributions of equivalent CQT expressions are equivalent.

This is trivially true by the definition of equivalence between E_1 and E_2. We know that there is a containment map between all predicates representing relations of equivalent CQT expressions. Moreover, because of safety, we know that the built-in predicates use only variables that are bound in (and hence captured by the containment mapping between) relational atoms.

Lemma 4.8 Source attributions of equivalent CQT expressions are equivalent.

Recall that source attribution is defined in terms of the free variables of a CQT expression. Because the queries are equivalent, we know that the two expressions define the same relation. Therefore, the containment mapping between relational predicates of equivalent expressions must take relational predicates containing free variables in E_1 to the corresponding relational predicates in E_2 and vice versa.

From Lemmas 4.7 and 4.8, we conclude that the comprehensive and source attributions of equivalent queries is equivalent. □

4.4.4 Multiple derivations from different expressions, composition

Composition, our reference for equivalent expressions defined on different databases, does not apply to non-predicate atoms, because theta-comparisons are not defined by expressions. We will, however, want to consider, the effect of non-predicate atoms on our definition for the attribution of composed expressions and whether the theorem for the recursive composition of attribution holds over theta-comparisons.

Again, we rely upon the fact that, though there is no unique, minimal query, there remains a containment mapping between the predicates in equivalent CQtheta queries.

What is the definition of a composed query (e.g. you can substitute expressions with the theta operator in it) and algorithm ... do you need to adjust either the drill down or the way you reconstruct the attribution as you back out?

Consequently, the introduction of inequality comparisons does not change the ability to compute attribution in a recursive fashion for predicates composed on other predicates.

Theorem 4.4 Composition holds for CQT expressions
Attribution composition computes the attribution of a composed CQT expression.

Assume without loss of generality the following CQT expressions E_1, E_2, E_3 defined by the formulas f, g, and h respectively s.t.
$E_1 \overset{\text{def}}{=} f = (p_1 \wedge p_2 \wedge \ldots \wedge p_n \wedge q)$ where q is the only IDB in E_1
$E_2 = q \overset{\text{def}}{=} g = (r_1 \wedge r_2 \wedge \ldots \wedge r_m)$ where $r_i \in d$
$E_3 \overset{\text{def}}{=} h = (p_1 \wedge p_2 \wedge \ldots \wedge p_n \wedge r_1 \wedge r_2 \wedge \ldots \wedge r_m)$
Again, we assume that variables in formula g of E_2 are renamed and reordered appropriately. The p's and r's may now include theta comparisons in addition to relational predicates with constants. We further assume, for convenience, that obvious redundancies are reduced (e.g. $(X < 10) \wedge (X < 5)$ reduces to simply $(X < 5)$)

Given E_1 defined on $d' = d \cup \{q\}$ and r, the result of evaluating E_1 on d', attribution composition computes the [comprehensive | source | relevant] attribution of result r in terms of d as defined by attr(r, E_3, d).

Lemma 4.9 $(a_3,h) \in A_3$ **is a comprehensive attribution for** E_3 **if and only if** $(a_3,h) \in$ **Compose** (A_1,f)**.**

This case is no different than for CQ expressions. That variables may now also appear in arithmetic comparisons does not affect their substitutions which are bound by the relational predicates.

Lemma 4.10 $(a_3,h) \in A_3$ **is a source attribution for** E_3 **if and only if** $(a_3,h) \in$ **Compose** (A_1,f)**.**

A_1 is the source attribution for E_1. Again, this follows the parallel for CQ expressions. Source attribution is defined by the relational predicates in which the free variables appear.

Lemma 4.11 $(a_3,h) \in A_3$ **is a relevant attribution for** E_3 **if and only if** $(a_3,h) \in$ **Compose** (A_1,f)**.**

A_1 is a relevant attribution for E_1. As with CQ expressions, we need to verify that $relevant(E_3) = relevant(E_1) + relevant(E_2)'$ where $relevant(E_2)' \subseteq relevant(E_2)$ and $relevant(E)$ refers to the relevant variables in E and likewise for $free(E)$; $bound(E)$. In other words, we want to verify that the relevant variables in E_3 are made up of the relevant variables in E_1 and the relevant variables in E_2. Because E_3 is the unification of E_1 and E_2, however, we avoid the problem observed in strict equivalence of identifying interactions between relational predicates and arithmetic comparisons. We form $relevant(E_2)'$ as we formed A_2' previously. We attribute only the relevant variables in q on the expression E_2.

We note that arithmetic comparisons may now constrain relational predicates containing free variables or relational predicates joined to predicates containing free variables. In addition, arithmetic comparisons may join relational predicates. However, comparisons in the r_j's of E_3 appear in E_2 and comparisons in the p_i's of E_3 appear in E_1. Furthermore, a comparison in the r_j's cannot include variables from the p_i's and vice versa, unless those variables appear in the IDB q of E_1. With these observations in mind, we proceed as in the case for CQ expressions.

Case(→)
Pick some relevant attribution $(a_3,h) \in A_3$ and split it: Project a_3 onto f and g.
We need to establish that $a_3(f) = (a_1,f) \in A_1$ and $a_3(g) = (a_2,g) \in A_2' \subseteq A_2$ where A_2' as the attribution for the tuple defined by $a_1 \cap a_2$, a tuple in q.

A substitution c/X is in a substitution list for a_3 because either X is free in E_3 or c/X joins two relational predicates, at least one of which is recursively joined to a relational predicate containing a free variable or is a constant from the original query expression that appears in a relational predicate recursively joined to a predicate containing a free variable of E_3.

Case 1. $X \in \mathit{relevant}(E_3)$ and $X \in \mathit{free}(E_3)$. $\mathit{Free}(E_3) = X \in \mathit{free}(E_1) \subseteq \mathit{relevant}(E_1)$ by definition of the equivalence of E_1 and E_3. For $X \in \mathit{free}(E_3) = Y \in \mathit{free}(E_2)$, Y must also be free in E_1 because E_2 is q in E_1 (e.g. $Y \in \mathit{free}(E_2) \rightarrow Y \in \mathit{free}(E_1)$). Consequently, at least for the relevant variables in E_2 that are free, we know $a_2 \in A_2' \subseteq A_2$

Case 2. $X \in \mathit{relevant}(E_3)$ joins relational predicates to a recursively joined set of relational predicates or X constrains one predicate in a recursively joined set of relational predicates (e.g. X is a constant in the formula or in an arithmetic comparison). All such predicates are in the set p_i and at least one joined predicate contains a free variable in h. Then X is relevant in E_1 so $a_3(f) = a_1$ for $(a_1,f) \in A_1$.

Case 3. $X \in \mathit{relevant}(E_3)$ is like Case 2 except all such predicates are in the set r_j. Then X is relevant in E_2 so $a_3(g) = (a_2,g) \in A_2' \subseteq A_2$ where A_2' as the attribution for the tuple defined by $a_1 \cap a_2$, a tuple in q. (recall that $g \overset{\text{def}}{=} q$ in E_1).

Case 4. $X \in \mathit{relevant}(E_3)$ appears in both some predicate p_i and some predicate r_j. Then X must appear in predicate q of E_1 (X is not free in E_1 so it must be bound in E_1 and appear in q in order to appear in both p_i and r_j in E_3). Therefore X is relevant in E_1 so $a_3(f) = a_1$ for $(a_1,f) \in A_1$.

It is possible that $a_3(g)$ is empty, which occurs when we consider a Cartesian Product and then do not restrict variables between arguments to the Cartesian Product. In this case, attribution relevance is trivially true in the p_i's.

From here, we do the same unfolding as before and conclude that **Compose** returns $a_1 \circ a_2(h)$ which is a_3.

Case (←)
Pick some random substitution list $a_1 \circ a_2$ as before and verify that $(a_1 \circ a_2,h) \in A_3$.
Proof by contradiction.
Suppose not. Then there must be a substitution $c/X \in a_1 \circ a_2$, $c/X \notin a_3$, or $c/Y \in a_3$, $c/Y \notin a_1 \circ a_2$.

Case 1. Pick Y. If Y is relevant in E_3, then Y must constrain the free variables in h in some way.
If Y is a free variable, then $c\ /Y \in a_1$, a contradiction.
If Y constrains a predicate containing a free variable through some recursively joined set of predicates amongst the p_i's, then $c/Y \in a_1$, a contradiction.
If Y constrains a predicate containing a free variable through some recursively joined set of predicates amongst the r_j's and is $a_1 \circ a_2$ as assumed above, then $c/Y \in a_2$, a contradiction.

If Y is relevant and appears in both some predicate p and some predicate r then $c/Y \in a_1$, a contradiction (see Case 4 for the ($\rightarrow$) direction).
Therefore, by contradiction, we conclude that there is no $c/Y \in a_3$, $c/Y \notin a_1 \circ a_2$.

Case 2. Pick X.
If $c/X \in a_1$ because it is a free variable in E_1, then by definition, $c/X \in a_3$, a contradiction.
If $c/X \in a_1$ because it constrains a free variable through predicates p_i, then $c/X \in a_3$, a contradiction.
If $c/X \in a_1$ and appears in both in q's and p's, then $c/X \in a_3$, a contradiction.
Now we need to be careful. Remember that a_2 is selected from attributing $a_1(g)$. It is possible for $a_1(g) = \{\}$ as in the case of Cartesian Product. attr($\{a_1(g)\}$, E_2, d') is non-empty only when there is a relevant variable in q.
If $c/X \in a_2$ because it is free in E_2 and free in E_1, then we know $c/X \in a_3$, a contradiction.
If $c/X \in a_2$ because it is free in E_2 and bound and relevant in E_1, then we know $c/X \in a_3$, a contradiction.
If $c/X \in a_2$ because it is bound in E_2, occurs among the predicates r_j and constraints a free variable in E_2 that is relevant in E_1 (through predicate q in E_1), then we know $c/X \in a_3$, a contradiction. □

Therefore, we conclude that attribution composition computes the attribution of a composed CQT expression. □

4.5 Adding explicit equality

Adding explicit equality to CQT expressions challenges our intuitions about the attribution of equivalent queries, but not necessarily in unexpected ways. The source of a variable is determined syntactically by occurrences of that variable in the expression. Logically, we say that the source of a variable is the predicate by which we limit (for purposes of safety) the values of a particular domain. Example 3.12 in Chapter 3 contrasted natural joins and explicit equality. We present a more abstract example here.

Example 4.25 Explicit equality

$$E_{20} = \{XY \mid p(XYZ) \wedge q(UVW) \quad (X = U)\}$$
$$E_{21} = \{XY \mid p(XYZ) \wedge q(XVW)\}$$
$$E_{22} = \{XZ \mid p(UWWX) \wedge (X = U)\}$$

In E_{21}, both predicates p and q may be said to limit values of X for purposes of safety. In E_{20}, predicate q does limit values of X, but only indirectly through an explicit comparison to U. In E_{22}, note that all variables are limited in the same predicate. More particularly, from the perspective of equivalence the examples introduce a slight irregularity into the containment map. We either implicitly push all equalities into the predicates (for example, eliminating variable U as in E_{21}) or rename all variables so that no variable name appears more than once as in E_{20}; all equalities are than explicit. Without the change, the containment map takes X and

U in E_{20} to X in E_{21}. Mapping from E_{21} to E_{20}, however is less clear. To what variable in E_{20} do we map E_{21} □

Rather than resolving the problem of explicit equality by either pushing equalities into relational predicates or renaming all variables, we suggest that the syntactic difference may prove useful for purposes of attribution. Under this interpretation, different relations that include the same domain may use the same domain variable to indicate multiple sources for that domain. In this way, we use the introduction of explicit equality to help differentiate attribution.

Example 4.26 Source attribution and explicit equality

E_{23} = {HNAME | regions(HNAME, REGION) ∧ sites("Imperial Palace", REGION) }
E_{24} = {HNAME | hostels(HNAME, PRICE, STATION) ∧ sites("Nakamise Dorsi", REGION) ∧
(REGION = STATION)}

E_{23}, adapted from Q2 in Chapter 3, attempts to locate hotels by the "Imperial Palace". It does so by matching the REGION in which the Imperial Palace is located, to the REGION in which individual hotels are located. Here, the two relations draw from the same domain so both relational predicates are considered sources for values of REGION.

E_{24}, adapted from Q5 in Chapter 3, attempts to locate hostels by "Nakamise Dorsi". However, the relation for hostels does not know about the domain of REGIONs. Rather, the query uses the knowledge that many train stations are named for the region in which they reside. As a consequence, we find hostels by equating values from the REGION domain with values from the STATION domain. □

Associating attribution with the syntax of a calculus expression allows us to distinguish between the concept of the natural join and the theta join (Ullman 1988). For purposes of attribution, in the natural join, two relations implicitly serve as sources for the same attribute domain. In theta-join, two attribute values, possibly from dissimilar domains, are explicitly compared.

Using the syntax of explicit equality to distinguish between different sources, however, clearly compromises the equivalence of comprehensive, source, and relevant attributions of strictly equivalent queries. We therefore offer

Observation 4.10 Attribution of strictly equivalent CQT$^+$ expressions

Though E_1 and E_2 are equivalent CQT$^+$ expressions (CQT expressions with explicit equality), then their [comprehensive | source | relevant] attributions, A_1 and A_2, are **not** necessarily equivalent. To see this, we need only recognize that the source attributions of equivalent expressions no longer necessarily map to one another as in E_{23} and E_{24} above. Comprehensive and relevant attribution suffer from the same issue. Although predicates map, there is not necessarily a consistent way of mapping domain variables between equivalent expressions. □

4.6 Adding union

In this next extension, we consider the addition of union into the query language. Unlike earlier extensions, union allows us to introduce and eliminate predicates from equivalent expressions. As a result, we first redefine our concept of attribution to account for union. We conclude that for the different types of attribution, the attribution of strictly equivalent queries are no longer necessarily equivalent. With some minor adjustments to the algorithm, however, we can show that attribution does continue to compose.

4.6.1 Attribution concept

Much as with the introduction of θ-comparison, the DRC imposes safety constraints on our introduction of disjunction in the language to support the semantics of algebraic union. In particular, the disjunction of two predicates must have the same set of arguments much as the algebraic condition on union requires union compatibility (Ullman 1988). As a further simplification for defining attribution in the presence of union, we assume prenex, disjunctive normal form (DNF) as the canonical form for all CQTU expressions. We know that we can transform a safe calculus expression into this form. A CQTU query therefore has the form:

$$\{X_1,\ldots,X_n \mid f_1(X_1,\ldots,X_n) \vee\ldots\vee f_m(X_1,\ldots X_n)\}$$

Every disjunct f_j is a CQT query that alone may make the expression true. In light of disjunction, we therefore generalize our original intuition for attribution. Attribute each disjunct as an independent CQT$^+$ query.

Definition 4.15 Attribution of the union of CQT$^+$ expressions
The [comprehensive | source | relevant] attribution of the disjunction of CQT$^+$ expressions is the union of the corresponding attributions for each constituent disjunct. □

Example 4.27 Attribution of the union of CQT$^+$ expressions

$E_{25} = \{A \mid p(ABC) \vee q(ABC)\}$

The [comprehensive | source | relevant] attribution for the expression is therefore the attribution of {A | p(ABC)} combined with the attribution of {A | q(ABC)} □

We actually saw several examples of unions from the examples in Chapter 3 beginning with Example 3.6. In the case of the union of CQ expressions, we know that there is a unique minimal equivalent (Ullman 1989). We find the unique minimal equivalent by minimizing each disjunct independently and then removing disjuncts that are contained by other disjuncts in the same expression. Under these limited circumstances, then, we can certainly argue that, for the unique minimal expression, the comprehensive, source, and relevant attributions are the same. For the general case of attribution equivalence of strictly equivalent queries, however, attribution equivalence breaks down with the introduction of union.

4.6.2 Multiple derivations – strict equivalence

For attribution, which is based upon substitutions, the problem posed by the introduction of union is immediately clear. Disjunction allows the introduction of new predicates, hence new variables and new substitutions. The containment condition for equivalent queries and the attendant mapping between attributions for equivalent expressions therefore breaks down. Example 3.7 of Chapter 3 offered one example of how attribution breaks down under union. Here, we consider a more abstract case. Consider the following equivalent expressions.

Example 4.28 Attribution of strictly equivalent expressions with disjunction

$E_{26} = \{A \mid p(ABC) \vee (p(ABC) \wedge q(ABC))\}$
$E_{27} = \{A \mid p(ABC) \vee (p(ABC) \wedge (C < 10))\}$
$E_{28} = \{A \mid p(ABC)\}$

The three queries are equivalent because the second disjunct in E_{26} and E_{27} is contained by the first disjunct. E_{26} and E_{27} therefore reduce to E_{26}. However, the comprehensive attribution for the first expression includes substitutions in the predicate q which do not map to the other equivalent expressions. Perhaps more obvious, we may regard q as a source for the attribute values of A in E_{26} although neither of the equivalent expressions reference q. For relevant attribution, we see that a variable, relevant in one disjunct, can prove irrelevant in a disjunct of the same expression or to an equivalent expression. In E_{27}, the attribute variable C is relevant in the second disjunct but neither in the first disjunct of the same expression nor in the third expression. □

That attribution breaks down under union corresponds to our intuitions about attribution. Attribution can provide corroborating information about the quality of a particular query result or the values in a particular result. Though redundant, attribution may also provide references to non-redundant ancillary information. Finally, from an intellectual property perspective, whether a source proves redundant or not, proper acknowledgement and perhaps remuneration is only appropriate.

4.6.3 Multiple derivations - composition

While attribution equivalence breaks down for strictly equivalent queries, we see that attribution continues to compose. As observed earlier for the relevant attribution of CQT expressions, composition assumes that we begin with a single formula and unfold the IDB. Composition does not reduce redundant disjuncts. We reason that we may unfold redundant disjuncts as easily as any other disjunct in the disjunction of CQT^{+} expressions (assuming also the appropriate renaming and reordering to avoid conflict in multiple occurrences of the same predicate or domain variable in the same disjunct).

We first update our algorithm to account for disjunctions. Then, we prove that the algorithm holds for the introduction of safe disjunction assuming that queries are expressed in canonical form.

To update the algorithm, we must first recall that the attribution of the expression is now the union of the attributions of each disjunct. We assume that the definition of any IDB may also include disjunction but that all IDB definitions are expressed in canonical form as well (i.e. the disjunction of conjuncts). The accumulation of disjuncts must therefore distribute in the original expression.

Example 4.29 Attribution of a composed expression with nested disjunction

$E_{29} = \{A \mid p(ABD) \vee q(ACE)\}$

$E_{30} \overset{def}{=} q(ACE) = \{ACE \mid (r(ABC) \wedge s(CDE)) \vee t(ACE)\}$

$E_{31} = \{A \mid p(ABD) \vee (r(ABC) \wedge s(CDE)) \vee t(ACE))$

E_{21} is an expression with an IDB in the second disjunct. The IDB, which we label E_{30}, itself contains a disjunction. Unifying the IDB gives E_{31}. Note the necessary variable renaming. Attribution is defined in terms of the base relations. As before, we want to discover whether we may iteratively attribute E_{29} and E_{30} in lieu of unifying the expression a priori.

Algorithm 4.2Attribution composition for CQT$^+$U expressions

Compose (A, s) where A is the attribution for s, a disjunction of CQT$^+$ sub-formulas, each of which may itself be a disjunction of CQT$^+$ sub-formulas.

```
Compose (A, s) {
if $s = \emptyset$ then return { }
else pick $f_i$ a disjunct in $s = f_1 \vee f_2 \vee \ldots f_x$
    $s := f_1 \vee f_2 \vee \ldots f_{i-1} \vee f_{i+1} \vee \ldots \vee f_x$
    $A' := \{(a,f) \mid (a,f) \in A \ and f = f_i\}$
    Compose (A, s) $\cup$ ComposeD $(A', f_l)$}

ComposeD $(A, f)$ {
if $f$ has no $q$'s then return $A$
else pick $q_i$, an IDB in $f$
    $f := p_1 \wedge p_2 \wedge \ldots \wedge q_{i-1} \wedge q_{i+1} \ldots q_m$
    ComposeD (Unfold $(A, q_i), f)$ }

Unfold $(A, q)$ {
if $A$ is $\emptyset$ then return { }
else pick $(a,f) \in A$
    let $g$ be the formula for IDB $E$ representing $q$
    let $u$ be the unifier for $h = unify(f,g)$
    let $E'$ be $E$ as defined by $g$ with the renaming of $u$
    $B$ = attr( $E'( a(q)/x )$, $d'$)
    Rewrite $(B, u(a - a(q)), h)$ $\cup$ Unfold $(A - \{(a,f)\}, q)$ }
Rewrite $(B, a, h)$ {
if $B$ is $\varnothing$ then return { }
else pick $(b,g) \in B$
    $\{<\{a \circ b\},h>\} \cup$ Rewrite $(B - \{(b,g)\}, a, h)$        □
```

This is the same algorithm as that presented for CQ expressions with the exception being lines (a) – (f). What was formerly called "Compose" we renamed "Compose Disjunct" or "ComposeD." As declared in line (a), "Compose" is now a function that recurses down the disjuncts in the formula for the query expression. We call "ComposeD" on each disjunct as if it were an isolated CQT$^+$ query. The attribution of the expression is then the union of the attributions from calling "ComposeD" on each disjunct. Because each substituion is defined for only one disjunct in the query expression, line (e) ensures that we ComposeD on each disjunct with only those substitutions applicable to a respective disjunct. We then propose:

Theorem 4.5 Attribution composition

Our algorithm for attribution composition computes the attribution for the union of composed CQT$^+$ expressions. Assume the following CQT+ expressions E_1, E_2, E_3 defined by the formulas f, g, and h respectively as:

$E_1 \overset{\text{def}}{=} f = (p_1 \wedge p_2 \wedge \ldots \wedge p_n \wedge q) \vee (t_1 \wedge t_2 \wedge \ldots)$ where q is the only IDB in E_1
$E_2 = q \overset{\text{def}}{=} g = (r_1 \wedge r_2 \wedge \ldots \wedge r_m) \vee (s_1 \wedge s_2 \wedge \ldots \wedge s_o)$ where $r_i, s_i \in d$
$E_3 \overset{\text{def}}{=} h = (p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m) \vee (p_1 \wedge \ldots \wedge p_n \wedge s_1 \wedge \ldots \wedge s_o) \vee (t_1 \wedge \ldots)$

Furthermore, we know that $(r_1 \wedge r_2 \wedge \ldots \wedge r_m)$ and $(s_1 \wedge s_2 \wedge \ldots \wedge s_o)$ are union compatible with schema defined by the IDB q.

Given E_1 defined on $d' = d \cup \{q\}$ and r, the result of evaluating E_1 on d', attribution composition computes the [comprehensive | source | relevant] attribution of result r in terms of d as defined by attr(r, E_3, d).

Lemma 4.12 $(a_3, h_i) \in A_3$ is a comprehensive attribution for E_3 if and only if $(a_3, h_i) \in$ Compose (A_1, f).

Case (→)

Pick a random substitution $(a_3, h_i) \in A_3$. Consider the following possibilities:

$h_i = (p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m)$
$h_i = (p_1 \wedge \ldots \wedge p_n \wedge s_1 \wedge \ldots \wedge s_o)$
$h_i = (t_1 \wedge \ldots)$

If $h_i = (t_1 \wedge \ldots)$ then we know that $(a_3,h_i) = (a_1,h_i) \in A_1$ because h_i is a disjunct in the formula for E_1 (see Algorithm 4.2 line (c)). For A' on $f_i = x_1 \ldots x_n = h_i$ we know that $I(h_i(a_3/x)) = true$ so $I(f_i(a_3/x)) = true$. There are no IDB in f_i so $(a_3,h_i) \in$ **ComposeD**$(A_1,f_i) \in$ **Compose**(A_1, f).

If $h_i = (p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m)$ in A_3 then we can say that $a_3(p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m)$, $(p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m) = a_1$, $(p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m) \in A_1$ because $\forall i$, $I(p_i(a_3(p_i)/x)) = true$ and $\forall j$, $I(r_j(a_3(r_j)/x)) = true$ and q is defined by the formula g. Or, to be more precise, $r_1 \wedge \ldots \wedge r_m$ is a disjunct of g that makes g true. Similarly, we know that $(a_3(r_1 \ldots r_m), g) = (a_2(r_1 \ldots r_m) \in A_2' \subseteq A_2$ (where A_2' is the attribution for tuple $a_1 \cap a_2$, a tuple in q). **Compose** calls **ComposeD** on $f_i = (p_1 \ldots q)$ with $A' = (a_1, f_i)$ in line (f) of Algorithm 4.2. **ComposeD** passes A' to **Unfold**. **Unfold** calls attr($\{a_3(q)\}$, E_2, d') which we already know is $A_2' \subseteq A_2$. **Unfold** is applied to every value of A' so certainly it calls itself on a_1 which we have already seen makes E_1 true. **Unfold** calls **Rewrite** with a_1 and A_2' so certainly it is applied to a_2. But **Rewrite** is called on h_i, the unification of $p_1 \ldots q$ and g and returns $(a_1 \circ a_2)$ which is a_3.

If $h_i = (p_1 \wedge \ldots \wedge p_n \wedge s_1 \wedge \ldots \wedge s_o)$ then we apply the same analysis as before, knowing that $h_i = (s_1 \wedge \ldots \wedge s_o)$ is a disjunct of g that also makes q true. As a consequence, it produces $A_2'' \subseteq A_2$ from attr($\{a_3(q)\}$, E_2, d') and we arrive at the same conclusion as before.

Case (←)
If (unify(f,g)) results in the disjuncts:

$(p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m)$,
$(p_1 \wedge \ldots \wedge p_n \wedge s_1 \wedge \ldots \wedge s_o)$ or
$(t_1 \wedge \ldots)$

does every $(a_1, t_1 \wedge \ldots)$ or $(a_1 \circ a_2, p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m)$ or $(a_1 \circ a_2, p_1 \wedge \ldots \wedge p_n \wedge s_1 \wedge \ldots \wedge s_o)$ appear as a substitution in A_3? Pick some arbitrary a_1 from a pair in A_1. If you picked some $(a_1, t_1 \wedge \ldots)$ then we know that a_1 makes disjunct $(t_1 \wedge \ldots)$ true. But because t is also a disjunct of E_3, if a_1 makes the disjunct true, then certainly it makes E_3 true therefore $(a_1, t_1 \wedge \ldots) \in A_3$.

Now if the pair $a_1 \in A_1$ is for disjunct $(p_1 \wedge p_2 \wedge \ldots \wedge p_n \wedge q)$ we want to pick an a_2 but not an arbitrary a_2. From Algorithm 4.2, **Compose** creates A_2' from attr($\{a_1(g)\}$, E_2, d'). So pick any a_2 from a pair $\in A_2'$. We know $a_1 \circ a_2$ paired with $(p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m) \vee (p_1 \wedge \ldots \wedge p_n \wedge s_1 \wedge \ldots \wedge s_o)$ appears in A_3 if it makes E_3 true. $\forall i$, $I(p_i(a_3(p_i)/x)) = true$ and $\forall j$, either $I(r_j(a_3(r_j)/x)) = true$ or $I(s_j(a_3(s_j)/x)) = true$. But is either disjunct true at the same time that the p's are true? Because we know that a_2 is from a pair in A_2' by construction, we know that a_2 makes E_2 true. Therefore, we know that $(a_1 \circ a_2, (p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m) \vee (p_1 \wedge \ldots \wedge p_n \wedge s_1 \wedge \ldots \wedge s_o)) \in A_3$. □

Lemma 4.13 **$(a_3,h_i) \in A_3$ is a source attribution for E_3 if and only if $(a_3,h_i) \in$ Compose (A_1,f) where A_1 is the source attribution for E_1.**

Case (→)
Pick a random substitution $(a_3,h_i) \in A_3$. Consider the following possibilities:

$h_i = (p_1 \wedge \dots \wedge p_n \wedge r_1 \wedge \dots \wedge r_m)$
$h_i = (p_1 \wedge \dots \wedge p_n \wedge s_1 \wedge \dots \wedge s_o)$
$h_i = (t_1 \wedge \dots)$

Regardless of which alternative is chosen, the source attribution consists of the free variables (and the accompanying variables in the associated relational predicate(s)).

If $h_i = (t_1 \wedge \dots)$ then we know then we know that $a_3(t_1 \wedge \dots)$, $(t_1 \wedge \dots) \in A_1$ because a_3 identifies the disjunct $(t_1 \wedge \dots)$ of E_3. But $(t_1 \wedge \dots)$ is also a disjunct of E_1, so this holds trivially. Note that there is no IDB in $(t_1 \wedge \dots)$ so Algorithm 4.2 line (2) returns the original source attribution for the disjunct $(t_1 \wedge \dots)$ for **Compose**$(A_1, (t_1 \wedge \dots))$.

If $h_i = (p_1 \wedge \dots \wedge p_n \wedge r_1 \wedge \dots \wedge r_m)$ then we observe that $a_3(p_1 \wedge \dots \wedge p_n \wedge q)$, $(p_1 \wedge \dots \wedge p_n \wedge q)$ is a pair $\in A_1$ because for predicates p_i, $E_1 \subseteq E_3$ and for predicate $q \in E_1$, q is defined in terms of the free variables of E_2 which is unfolded in a_3. Similarly, we can say that $(a_3(r_1 \wedge \dots \wedge r_m))$, $q = (a_2,g) \in A_2' \subseteq A_2$ where A_2' is the source attribution of $a_1 \cap a_2$, a tuple of q. **Compose** passes $f' = (p_1 \wedge \dots \wedge p_n \wedge q)$ to **ComposeD** with source attribution A' defined in terms of the p's and q's. **ComposeD** passes A' to **Unfold** with $q = (r_1 \wedge r_2 \wedge \dots \wedge r_m) \vee (s_1 \wedge s_2 \wedge \dots \wedge s_o)$. **Unfold** is called on every source substitution in A' so certainly it is called on a_1. **Unfold** calls for attr($\{a_1(g)\}$, E_2, d') which we know includes the source substitutions $A_2' \subseteq A_2$ where the formula in the attribution pair is the disjunct $(r_1 \wedge r_2 \wedge \dots \wedge r_m)$. **Rewrite** is called on every element of A_2' so eventually it is called on a_2. But **Rewrite** pairs $a_1 \circ a_2$ with h_i a disjunct $(p_1 \wedge \dots \wedge p_n \wedge r_1 \wedge \dots \wedge r_m)$ of unify(f,g) in line (10) of Algorithm 4.2. This, then, is just a_3. The same reasoning applies for the disjunct $(s_1 \wedge s_2 \wedge \dots \wedge s_o)$ from the attribution in **Unfold**.

Case (←)
If (unify(f,g)) results in the disjuncts:

$(p_1 \wedge \dots \wedge p_n \wedge r_1 \wedge \dots \wedge r_m)$,
$(p_1 \wedge \dots \wedge p_n \wedge s_1 \wedge \dots \wedge s_o)$ or
$(t_1 \wedge \dots)$

does every $(a_1, t_1 \wedge \dots)$ or $(a_1 \circ a_2, p_1 \wedge \dots \wedge p_n \wedge r_1 \wedge \dots \wedge r_m)$ or $(a_1 \circ a_2, p_1 \wedge \dots \wedge p_n \wedge s_1 \wedge \dots \wedge s_o)$ appear as a substitution in A_3? For pairs $(a_1, t_1 \wedge \dots)$ then we know that a_1 is a source substitution $(t_1 \wedge \dots)$. But because t is also a disjunct of E_3, if a_1 is a valid source substitution for E_1, then certainly it is likewise for E_3 therefore $(a_1, t_1 \wedge \dots) \in A_3$.

For pairs involving an $a_1 \circ a_2$ pick some arbitrary a_1 from a pair in A_1. Now pick an a_2 from $A_2' \subseteq A_2$ generated by the attr($\{a_1(g), E_2, d'\}$) in **Unfold**. This will give a substitution a_2 either in $(r_1 \wedge r_2 \wedge \ldots \wedge r_m)$ or $(s_1 \wedge s_2 \wedge \ldots \wedge s_o)$. We know that $a_1 \circ a_2$ paired with $(p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m)$ or $(p_1 \wedge \ldots \wedge p_n \wedge s_1 \wedge \ldots \wedge s_o)$ makes A_3 true if it makes E_3 true. And we know a_1 makes the p's true just as a_2 makes the r's or the s's true by construction. Therefore, we know $(a_1 \circ a_2, (p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m)) \in A_3$ and $(a_1 \circ a_2, (p_1 \wedge \ldots \wedge p_n \wedge s_1 \wedge \ldots \wedge s_o)) \in A_3$ □

Lemma 4.14 $(a_3,h) \in A_3$ is a relevant attribution for E_3 if and only if $(a_3,h) \in$ Compose (A_1,f).

Where A_1 is a relevant attribution for E_1. As in prior cases, we need to verify that *relevant*(E_3) = *relevant*(E_1) + *relevant*$(E_2)'$ where *relevant*$(E_2)' \subseteq$ *relevant*(E_2) and *relevant*(E) refers to the relevant variables in E and likewise for *free*(E); *bound*(E). We form *relevant*$(E_2)'$ as we formed A_2' previously. We attribute only the relevant variables in q on the expression E_2. With disjunction, there is the additional complexity of tracking relevance in each disjunct.

Case (→)

Pick a random substitution $(a_3,h_i) \in A_3$. Consider the following possibilities:

$h_i = (p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m)$
$h_i = (p_1 \wedge \ldots \wedge p_n \wedge s_1 \wedge \ldots \wedge s_o)$
$h_i = (t_1 \wedge \ldots)$

Suppose $h_i = (t_1 \wedge \ldots)$. We also know that $(t_1 \wedge \ldots)$ is a disjunct of E_1 which means that $(a_3, (t_1 \wedge \ldots)) \in A_1' \subseteq A_1$. Because there are no IDB in this disjunct, we know that the call to **ComposeD** on $(t_1 \wedge \ldots)$ with $A_1' \subseteq A$ for pairs $(a_3, (t_1 \wedge \ldots))$ simply returns A_1'. So we conclude $(a_3, (t_1 \wedge \ldots)) \in$ **Compose**$(A_1, (t_1 \wedge \ldots))$.

If $h_i = (p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m)$ then we consider the same cases as for relevance in CQ. However, we must now consider the cases for each disjunct.

Case 1. $X \in$ *relevant*$(p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m)$ and $X \in$ *free*$(p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m)$. We know that *free*$(p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m) = X \in$ *free*$(E_1) \subseteq$ *relevant*(E_1) by definition of the equivalence of $p_1 \wedge \ldots \wedge p_n \wedge q$ and $h_i = p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m$. Consequently, at least for relevant variables in the disjunct $(p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m)$ that are free, we know $a_2 \in A_2' \subseteq A_2$.

Case 2. $X \in$ *relevant*$(p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m)$ joins relational predicates to a recursively joined set of relational predicates or X constrains one predicate in a recursively joined set of relational predicates (e.g. X is a constant or X appears multiple times in a single relation). All such predicates are in the set p_i and at least one joined predicate contains a free variable in $h_i = p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m$. Then X is relevant in E_1 so $a_3(f) = a_1$ for $(a_1, p_1 \wedge \ldots \wedge p_n \wedge q) \in A_1$.

Case 3. $X \in relevant(p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m)$ as in Case 3 of Lemma 4.6 where the X's appear only in the r_j's. Then $X \in relevant(r_1 \wedge \ldots \wedge r_m)$ which implies $X \in relevant(E_2)$ so $a_3(g) = (a_2, g) \in A_2' \subseteq A_2$ where A_2' as the attribution for the tuple defined by $a_1 \cap a_2$, a tuple in q. Of course A_2' may also include some substitutions from other disjuncts in the definition of q (e.g. $s_1 \wedge \ldots \wedge s_o$).

Case 4. $X \in relevant(p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m)$ appears in both some predicate p_i and some predicate r_j. Then X must appear in predicate q of E_1 (X is not free in E_1 so it must be bound in E_1 and appear in q in order to appear in both p_i and r_j in E_3). Therefore X is relevant in E_1 so $a_3(f) = a_1$ for $(a_1, p_1 \wedge \ldots \wedge p_n \wedge q) \in A_1$.

It is possible that $a_3(g)$ is empty, which occurs when we consider a Cartesian Product and then do not restrict variables between arguments to the Cartesian Product (i.e. no θ comparisons). In this case, attribution relevance is trivially true in the p_i's.

From here, we do the same unfolding as before and conclude that **Compose** returns $(a_1 \circ a_2, (p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m)) \in A_3$. We can do the same analysis for $h_i = p_1 \wedge \ldots \wedge p_n \wedge s_1 \wedge \ldots \wedge s_o$ or any other disjunct of q.

Case ($\leftarrow$)
Pick some random substitution from **Compose**: $(a_1, (t_1 \wedge \ldots))$ or $(a_1 \circ a_2, f')$ where f' is a disjunction $(p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m)$ or $(p_1 \wedge \ldots \wedge p_n \wedge s_1 \wedge \ldots \wedge s_o)$ and verify that it appears in A_3.

Proof by contradiction.
Suppose not. Then there must be a substitution:

$c/X \in a_1$ where $(a_1, (t_1 \wedge \ldots)) \notin A_3$ or some
$c/X \in a_1 \circ a_2$ where $(a_1 \circ a_2, p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m) \notin A_3$ or some
$c/X \in a_1 \circ a_2$ where $(a_1 \circ a_2, p_1 \wedge \ldots \wedge p_n \wedge s_1 \wedge \ldots \wedge s_o) \notin A_3$.

But we know $(t_1 \wedge \ldots)$ is a disjunct in E_3 so if a_1 is relevant in the t's for E_1 then it must still be relevant in the same disjunct of E_3. A contradiction.

If $c/X \in a_1$ because it is free in E_1 then by definition, $c/X \in a_3$, a contradiction.
If $c/X \in a_1$ because it constrains a free variable through the p's then $c/X \in a_3$.
If c/X appears in both the p's and predicate q, then $c/X \in a_3$ by definition.

Now we need to be careful. Remember that a_2 is selected from attributing $a_1(g)$. It is possible for $a_1(g) = \{\}$ as in the case of Cartesian Product. attr($\{a_1(g)\}$, E_2, d') is non-empty only when there is a relevant variable in q. Recall that E_2 is in DNF so the free variables are the same in each disjunct of E_2.

If $c/X \in a_2$ because it is free in E_2 and free in E_1, then we know $c/X \in a_3$, a contradiction.

If $c/X \in a_2$ because it is free in E_2 and bound and relevant in E_1, then we know $c/X \in a_3$, a contradiction.

If $c/X \in a_2$ because it is bound in E_2, occurs among the predicates r_j of a disjunct in E_2 (similarly for the other disjuncts of E_2 i.e. s_j) and constraints a free variable in E_2 that is relevant in E_1 (through predicate q in E_1), then we know $c/X \in a_3$, a contradiction. □

Therefore, we conclude that attribution composition computes the attribution of a composed expression. □

By representing our expressions in DNF, we can treat each disjunct independently and compose in a depth first manner across all disjuncts and all IDB. As before, we can easily imagine unfolding successive levels of IDB.

4.7 Adding negation

Negation, in general, poses problems for query evaluation (Abiteboul, Hull, and Vianu 1995). Likewise, negation presents problems for attribution. From Chapter 3, the intuition behind attribution for negation corresponds to the logical interpretation of safe expression. We can confirm the truth of a negated assertion (fact in the database) by verifying that the (positive) assertion itself does not exist in the database. Unfortunately, this intuition breaks down under composition of queries with negation. We identify a subset of queries with negation under which composition is preserved.

4.7.1 Attribution concept

As indicated in Chapter 3, to verify that the (positive) assertion does not exist, the attribution must therefore consider (include) every true substitution for the negated sub-formula. We first illustrated this intuition in Example 3.13 and Example 3.14 of Chapter 3.

Example 4.30 Attribution for an expression with negation

$E_{32} = \{ABC \mid r(ABC) \wedge \neg s(ABC)\}$

To verify that a substitution <1/A, 2/B, 3/C> is in the attribution for the expression, we must not only verify that I(r(1/A, 2/B, 3/C)) = true but also that for every substitution <x/A, y/B, z/C> such that I(s(x/A, y/B, z/C)) = true, x ≠ 1 or y ≠ 2 or z ≠ 3. □

Moreover, if there is more than one negated predicate, we need to confirm that a valid substitution for the expression does not make any of the negated predicates true. We would do so by confirming that a substitution for the formula does not include any true substitution for any negated predicate.

4.7.2 Types of attribution

To formalize attribution in the context of negation, we introduce a few additional assumptions. First, using standard rules, all negations are pushed down to the level of individual predicates. The negation of an arithmetic comparison is simply expressed as its logical converse (e.g. $\neg (X < Y) \equiv (X \geq Y)$). Second, formulas continue to be flattened as the disjunction of conjuncts where all conjuncts are either positive or negative predicates or theta comparisons. Within each disjunct, negated predicates are limited for safety as per the syntactic rules described earlier. A formula is therefore a disjunction of conjuncts of the form:

$$p_1 \wedge p_2 \wedge \ldots \wedge p_n \wedge \neg q_1 \wedge \neg q_2 \wedge \ldots \wedge \neg q_m \wedge t_1 \wedge \ldots \wedge t_o$$

where the p's are non-negated predicates, the q's are negated predicates, and the t's are theta comparisons. For safety, for each j in m, every argument in q_j must also appear in some predicate p_i or bound to a constant. Based upon these extensions to address negation, we can now redefine what we mean by attribution.

Definition 4.16 Comprehensive attribution
The comprehensive attribution for an expression in DNF, possibly with negated predicates, is the union of the comprehensive attributions for each disjunct, f. The comprehensive attribution for each disjunct is a set of triples $<a, n, f>$ where a is a substitution for which the non-negated predicates p_i and θ-comparisons t_o in disjunct f evaluate to true and n is itself a set of substitutions $\{<b, m, q_j>\}$. The set n ranges over all of the negated predicates q_j and includes every substitution b that makes q_j true. Assuming that there is no b that agrees in the corresponding substitutions for values of a $I(q_j(a)) = false$ we may then concludes $I(\neg q_j(a)) = true$. By default, m is $\varnothing$. □

In source attribution, the intuition is that we want to know the predicates (and their corresponding substitutions) from which values in the query result are drawn. Therefore, only non-negated predicates are considered as possible sources. Negated predicates because they do not match our intuition as a source for values in the result.

Definition 4.17 Source attribution
The source attribution for an expression in DNF, possibly with negated predicates, is the union of the source attributions for each disjunct, f. The source attribution for each disjunct is a set of triples $<a', n, f>$ where a' is a sublist of substitutions a for non-negated predicates of f that contain free variables and make f true. n is $\varnothing$. □

For relevant attribution we want to consider variables that in some way affect the result. Because of the safety requirement, renaming any variable in a negated predicate would compromise the expression. As a consequence, any variable in a negated predicate is relevant and we have the same issue as introduced in comprehensive attribution for capturing all appropriate substitutions.

Definition 4.18 Relevant attribution
The comprehensive attribution for an expression in DNF, possibly with negated predicates, is the union of the relevant attributions for each disjunct, *f*. The relevant attribution for each disjunct is a set of triples $<a, n, f>$ where a is a substitution for all relevant variables in *f* that make *f* true. All variables in the head (free in the formula for the expression) are relevant. In addition, a bound variable is relevant to the result if renaming the variable to some name not already in the expression (or eliminating a constant) would relax a constraint on one or more of the attribute domains in the result relation (free in the formula for the expression). By definition, any variable in a negated predicate is relevant. Therefore, as with comprehensive attribution, n is itself a set of substitutions $\{<b, m, q_j>\}$. The set n ranges over all of the negated predicates q_j and includes every substitution b that makes q_j true. We therefore know that $I(q_j(a)) = false$ and $I(\neg\, q_j(a)) = true$. By default, m is $\varnothing$. □

4.7.3 Attribution equivalence and composition

Having updated our definition of attribution, we consider the impact of introducing negation on our attribution properties. Determining the equivalence of queries with negation is an open question that has persisted for many years (Abiteboul, Hull, and Vianu 1995). It is not an issue that we will attempt to resolve here. Consequently, claims about the attribution of equivalent queries with negation are also outside the scope of this thesis.

However, as seen in our discussion of attribution for CQT expressions, we can address the issue of attribution composition separately. With the introduction of negation, it is apparent that, in general, the property of composition no longer holds. We cannot calculate the attribution for a query result by recursively tracing backwards through each sub-formula. However, we identify a subset of queries under which composition continues to hold.

First, we notice that, in the general case, nested negations (i.e. $b \equiv \neg(\neg b)$) compromises our ability to compose attribution.

Example 4.31 Intersection of predicates a and b using nested negation
Consider two expressions E_{33} and E_{34} with the following formulas.
$f_{33} = a \wedge \neg\, (a \wedge \neg\, b)$
$f_{34} = b \wedge \neg\, (b \wedge \neg\, a)$

Logically, we know that $E_{33} \equiv E_{34}$. Indeed when we put E_{33} and E_{34} into canonical form by pushing and distributing the negation, we end up with $f_{33} = f_{34} = a \wedge b$. However, suppose we defined the following IDB:

$c = a \wedge \neg b$
$d = b \wedge \neg a$

A substitution in the attribution of c includes values for variables in a and every substitution that makes b true. Likewise for a substitution in the attribution of d. Consider again our original expressions now defined using IDB c and d.

$f_{33}' = a \wedge \neg c$
$f_{34}' = b \wedge \neg d$

By expanding c and d and pushing down the negations, we know that the source attribution for E_{33} = source attribution for E_{34} = source attribution for $(a \wedge b)$. However, we can equally see that the source attribution for E_{33}' = substitutions in A while the source attribution for E_{34}' = substitutions in b.

Similarly, negations are fully eliminated in the canonical form of E_{33} and E_{34} suggesting that a comprehensive or relevant substitution in the attribution for these expressions will be a single list of variables that make a and b true. However, E_{33}' and E_{34}' contain negated literals suggesting that a substitution will include a list of variables that make a (or b respectively) true and then a set of all substitutions that make c (or d respectively) true. Composition would then recurse on all substitutions in c (or d) rather than a single substitution as in $(a \wedge b)$. □

We can think of the phenomenon in the example above as an additivity property that reflects attribution composition. If R is an expression composed on Q and *r* is a result in both Q and R, then the attribution for *r* in R should at least include the substitutions for the attribution of *r* in Q. Unfortunately, as seen in the example above, composition breaks down when we allow negations to cancel one another.

The problem extends beyond nested negations, however. As demonstrated below, distributing negation over conjunction also violates the additivity observed above.

Example 4.32 Distributing negation over conjunction
Imagine expressions with the following formulas.

$f_{35} = C \wedge \neg (A \wedge B)$
$f_{36} = (C \wedge \neg A) \vee (C \wedge \neg B)$

Here, we see that the attribution for the first is not the same as the attribution for the second because of what you associate in the attribution. Logically the two are equivalent. However, a triple in the first expression has $n = \{b, m, (A \wedge B) \mid I(b/X(A \wedge B)) = true\}$. A triple in the second expression looks like either $\{b, m, (A) \mid I(b/X(A)) = true\}$ or $\{b, m, (B) \mid I(b/X(B)) = true\}$. It is straightforward to see that for substitutions (a, n, f) where a is only absent from A

or from B but not both, that the substitutions could look quite different. as a consequence, it is clear that negation poses some problems for our intuitions about attribution.
□

However, by further constraining the syntactic rules under which we may negate predicates, we arrive at a rudimentary subset of the DRC where negation is permitted yet attribution composition is preserved.

Definition 4.19 Attributable expression.
To define an attributable expression, we extend the rules for safety presented at the beginning of this Chapter (Ullman 1988). In particular, we introduce the concept of a negatable formula. Only a negatable formula may be negated and remain attributable.

1. Any atom is a formula and is negatable.
2. The disjunction of non-negated atoms is a negatable sub-formula.
3. The disjunction of negatable sub-formulas is negatable. □

Examples 4.33 Negatable sub-formulas in the safe DRC
$f_{37} = A \wedge \neg B \wedge \neg C$
Where A, B, and C are relational predicates representing base relations. Note that the rules of safety require that every variable appearing in B and C also appear in A.

$f_{38} = A \wedge (B \vee C)$
The $(B \vee C)$ is a negatable sub-formula. When we push the negation into the formula, then the formula becomes the same as the first formula.

$f_{39} = (A \vee B) \vee (C \vee D)$ is a disjunction of negatable sub-formulas that are negatable on their face. However, were either of the expressions already negated, then the formula would no longer be negatable. □

We suggest that the attribution of attributable expressions composes. Because we have updated our definitions of attribution to account for negation, our algorithm for composing attributions requires corresponding updates. We first amend our algorithm for calculating attribution and then prove that, for negatable expressions, that the algorithm calculates the attribution for an extended expression.

Algorithm 4.3Attribution composition for negatable query expressions

Compose (A, s) where A is the attribution for *s*, a disjunction of CQT^+ sub-formulas with negated predicates, each of which may itself be a disjunction of CQT^+ sub-formulas with negated predicates.

FORMAL MODEL

```
Compose (A, s) {                                                   (a)
if $s = \varnothing$ then return { }                                           (b)
else pick $f_i$ a disjunct in $s = f_1 \vee f_2 \vee \ldots f_x$                     (c)
    $s := f_1 \vee f_2 \vee \ldots f_{i-1} \vee f_{i+1} \vee \ldots \vee f_x$                  (d)
    $A' := \{(a,f) \mid (a,f) \in A\ and\ f = f_i\}$                         (e)
    Compose (A, s) $\cup$ ComposeD ($A'$,$f_1$)}                          (f)

ComposeD ($A$, $f$) {                                                (1)
if $f$ has no $q$'s then return $A$                                       (2)
else pick $q_i$, an IDB in $f$                                           (3)
    $f := p_1 \wedge p_2 \wedge \ldots \wedge q_{i-1} \wedge q_{i+1} \ldots q_m$                         (4)
    if $q_i$ is negated                                                (5)
    then ComposeD (UnfoldN ($A$, $q_i$), $f$)                              (6)
    else ComposeD (Unfold ($A$, $q_i$), $f$) }                              (7)

UnfoldN ($A$, $q_i$) {                                                 (8)
if $A$ is $\varnothing$ then return { }                                          (9)
else pick some triple $<a, n, f> \in A$                                  (10)
    let $g$ be the formula for the definition of $q_i$                    (11)
    let $u$ be the unifier for $h = unify(f,g)$                          (12)
    $n' :=$ RewriteN ($n$, $u$, $q_i$)                                       (13)
    $\{<u(a), n', h>\}$ $\cup$ Unfold ($A - <a, n, f>$, $q_i$)}                  (14)

RewriteN ($n$, $u$, $q_i$) {                                              (15)
foreach triple $<b, \varnothing, q>$ in $n$ where $q = q_i$                          (16)
    $n := n - \{b, \varnothing, q\}$                                           (17)
let $g$ be the formula for the definition of $q_i$                        (18)
$B$ = attr($u(g)$, $d'$)                                                  (19)
$n := n \cup B$}                                                        (20)

Unfold ($A$, $q$) {                                                     (21)
if $A$ is $\varnothing$ then return { }                                         (22)
else pick $(a, n, f) \in A$                                              (23)
    let $g$ be the formula for IDB $E$ representing $q$                   (24)
    let $u$ be the unifier for $h = unify(f,g)$                          (25)
    let $E'$ be $E$ as defined by $g$ with the renaming of $u$             (26)
    $B$ = attr( $E'$( $a(q)/x$ ), $d'$)                                       (27)
    Rewrite ($B$, $u(a - a(q))$, $h$) $\cup$ Unfold ($A - \{(a,f)\}$, $q$) }       (28)

Rewrite ($B$, $a$, $h$) {                                               (29)
if $B$ is $\varnothing$ then return { }                                         (30)
```

```
else pick $(b, m, g) \in B$                                    (31)
    $\{<\{a \circ b\}, n \cup m, h>\} \cup$ Rewrite $(B - \{(b,g)\}, a, h)$        (32)        □
```

We took our original algorithm and first extended it to account for unions. Here, we make several changes to account for negation. First and foremost, we extended attribution from a pair to a triple consisting of a substitution list (a), a formula (f) to which the substitution list provides a true interpretation, and a set consisting of the attributions for each negated predicate in the formula. As a consequence, the descendants of our initial functions to unfold and rewrite are updated to return triples in lines (23), (29), and (32). More significantly, we must now consider IDB whose definition includes negated predicates as well as negated IDB.

We calculate the attribution of an IDB with negated predicates in line (27). We know that for attributable expressions, the unification of our original formula with the definition of the IDB in line (25) simply adds additional, negated conjuncts. Consequently, we may simply combine attributions for negated predicates in the original expression with attributions for negated predicates in the IDB as seen in line (32).

For negated IDB that are also attributable, we know that certain conditions must hold. Specifically, we know that the IDB must be a disjunction of non-negated predicates. Pushing negations down, this translates into a unifier that effectively substitutes a conjunction of negated predicates for one negated predicate. Accordingly, for each attribution triple of the original formula, we simply remove the attributions for the negated IDB. This is done in lines (15) – (17). In place of these attributions, we substitute the attributions for each predicate in the definition of the IDB. Note that in line (19), we simply attribute the formula for the IDB (assuming the unifier u to avoid conflicts in variable naming). If the IDB is a disjunction, then the attribution will comprise the union of the attributions for each disjunct.

Based upon this revised algorithm, we now offer:

Theorem 4.6 Attribution composition
Our algorithm for attribution composition computes the attribution for attributable expressions.

For IDB that do not include negation, the algorithm is unchanged except for the introduction of a third component to the substitution (which is empty in the case of no negated predicates). Under this circumstances, the proof therefore follows that of Theorem 1.5. More interesting are the two cases of IDB that include negations and negated IDB.

Our algorithm for attribution composition computes the attribution for the union of attributable, composed CQT^{+} expressions. Assume the following CQT+ expressions E_1, E_2, E_3 defined by the formulas f, g, and h respectively as:

$E_1 \overset{\text{def}}{=} f = (p_1 \wedge p_2 \wedge \ldots \wedge p_n \wedge q) \vee (t_1 \wedge t_2 \wedge \ldots)$ where q is the only IDB in E_1

$E_2 = q \overset{\text{def}}{=} g = (r_1 \wedge r_2 \wedge \ldots \wedge r_m) \vee (s_1 \wedge s_2 \wedge \ldots \wedge s_o)$ where r_i, $s_i \in d$
$E_3 \overset{\text{def}}{=} h = (p_1 \wedge \ldots \wedge p_n \wedge r_1 \wedge \ldots \wedge r_m) \vee (p_1 \wedge \ldots \wedge p_n \wedge s_1 \wedge \ldots \wedge s_o) \vee (t_1 \wedge \ldots\ )$

Note that subject to safety, any of the predicates (with the exception of the IDB q) may be negated. To negate the IDB q, as articulated in Definition 1.17, we are limited to disjunctions of non-negated predicates. Our IDB are thus limited to expressions of the form: $E_4 = q \overset{\text{def}}{=} g = r_1 \vee r_2 \vee \ldots \vee r_m$

Given E_1 defined on $d' = d \cup \{q\}$ and r, the result of evaluating E_1 on d', attribution composition computes the [comprehensive | source | relevant] attribution of result r in terms of d as defined by attr(r, E_3, d).

Lemma 4.15 ***$(a_3,h_i) \in A_3$*** **is a comprehensive attribution for E_3 if and only if** ***$(a_3,h_i) \in$*** **Compose *(A_1,f)*.**

Case ($\rightarrow$)
We first consider the case where the IDB itself is not negated although any of the base relational predicates (e.g. $r \in d$) may be negated (subject to safety). Pick some $(a_3, n_3, h) \in A_3$. We know that $a_3(f)$ and $a_3(g)$ provide substitutions for the non-negated predicates in a disjunct of f and g by definition. Furthermore, we know that $n_3 = \{(b_1,m,k) | m = \varnothing \wedge k$ is a negated predicate in a disjunct of h that appears also in the corresponding disjunct of $f\} \cup \{(b_2,m,k) | m = \varnothing \wedge k$ is a negated predicate in a disjunct of h that appears also in the corresponding disjunct of $g\}$. For every negated predicate $\neg k$, we know that n_3 includes every substitution b that makes the non-negated predicate k true whether the predicate is in E_1 or E_2. Thus we can model the proof for Lemma 4.10 to verify that the property holds for non-negated predicates and we know that the property holds for n_3, the set of substitutions for negated predicates in h_i.

What then if we allow the IDB q to be negated? We know that to be attributable, the IDB must be defined in the form of E_4, a disjunction of attributable subformulas. Second, we know that when unfolded, pushing down the negation transforms the disjunction into a conjunction where every predicate in g (the formula for E_4) is negated. So in h, by definition for the attribution of negated predicates, n_3 includes the union of the set of all true substitutions for each negated conjunct. But we know that (b_1,m,k) when k is an IDB in E_1 will include every true substitution b_1 for the negated predicate k. Moreover, the negated IDB q is safe in E_1 but were we to attempt attributing the negation of the formula for E_2, we would have an unsafe expression. Instead, we know that q is negated so we call **UnfoldN** instead. In the subsequent call to **RewriteN** we see how we remove the positive substitutions for q (See Algorithm 4.3 line (17)) and replace substitutions in q with the full set of substitutions that make the $u(g)$ (the formula for E_4 subject to appropriate renaming) true (See Algorithm 4.3 line (19)). Thus, we see that n_3 is again $n_1 \cup n_2$ (minus the substitutions for q which do not appear in h_i) and we conclude that ($\rightarrow$) holds.

Case (←)
Suppose now that you have some $(a_1 \circ a_2, n_1 \oplus n_2, h_i)$ where h_i is a disjunct of h and $\oplus$ denotes the union of n_1 and n_2 subject to the removal of substitutions for the IDB of n_1 that are unfolded in n_2. (Note that if the IDB is not negated, then $\oplus$ reduces to $\cup$). We pick some $a_1 \in A_1$ and pick a_2 by construction as before. Now, we know that $a_1 \circ a_2$ gives substitutions for non-negated predicates in a_3 as before. However, a_2 now also includes $\{(b_2,m,k)\}$ for negated predicates in g likewise for $\{(b_1,m,k)\}$ in f. But each negated predicate in f and each negated predicate in g is also negated in h by our limitation on attributable expressions. As a consequence, we know that we can $a_1 \circ a_2$ gives a_3 and that $n_1 \oplus n_2$. Hence, we may conclude that (←) holds. □

Lemma 4.16 $(a_3,h_i) \in A_3$ is a source attribution for E_3 if and only if $(a_3,h_i) \in$ Compose (A_1,f) where A_1 is the source attribution for E_1.

We know by definition that a source attribution does not include substitutions in negated predicates. Therefore, we need only consider the case where predicates other than the IDB are negated. Therefore, we only unfold non-negated IDB and consider only source substitutions in non-negated predicates. We see from Lemma 4.13 that the substitutions in both the negated and non-negated predicates compose. Thus, we conclude that the proof then mirrors the proof for the composition of source attributions for the union of CQT$^+$ subformulas in Lemma 4.11. In particular, note that ruling out negated IDB, a negated predicate in f or g corresponds to a negated predicate in h and vice versa. Likewise for non-negated predicates. See Lemma 4.11 for the case of a free variable in the IDB. □

Lemma 4.17 $(a_3,h) \in A_3$ is a relevant attribution for E_3 if and only if $(a_3,h) \in$ Compose (A_1,f).

Where A_1 is a relevant attribution for E_1. The challenge in prior classes of queries was to verify that variables relevant in E_1 and E_2 respectively were relevant in E_3 and vice versa. In this way, we could construct relevance in the iterative manner of comprehensive and source attribution. For variables in negated predicates, however, this is trivially true simply because any negated domain variable is defined as relevant. Consider negated predicates in f or g (apart from the IDB). Then, the same variables and predicates are relevant in the unfolding to h and thus relevant. For a negated IDB, our condition on attributable expressions guarantees that every predicate in g is a negated conjunct and is therefore relevant. Thus, for negated IDB, we simply substitute the negated IDB in (a_1, n, f) with every positive substitution in n. Furthermore, for purposes of safety, every variable in the IDB (q) of f must be relevant in the non-negated predicates and so must also appear in h. □

Thus, building heavily upon Theorem 4.5, we conclude that attribution composition computes the attribution of a composed expression provided that the constituent expressions are attributable. □

4.8 Summary

We began with an overview of the domain relational calculus upon which we build our formal attribution model. We first define our attribution model for simple, conjunctive queries. The model includes definitions for three different types of attribution as well as several different properties of these different attribution types. In particular, we use the properties of conjunctive queries to identify three different categories of equivalence properties and granularity principles.

Having presented a preliminary model, we generalize the model by progressively increasing the expressiveness of the query language for which the model is defined. In the first step, we introduce arithmetic comparisons (omitting explicit equality). Our reliance upon conjunctive query properties to establish equivalence causes conclusions about "relevant" attribution to break down under theta operators. We indicate how explicit equality compromises the attribution of strictly equivalent query expressions.

Subsequent steps introduce union and then negation into the query model. Composition is the only property that continues to hold when unions are permitted. Finally, all attribution properties fail upon incorporation of negation into the query language. However, we define a subset of *attributable expressions* for which the property of composition is preserved.

5 Extended algebra

Unfortunately, while practical systems today are rooted in the Domain Relational Calculus from which we draw our definitions for attribution, conventional systems do not query using the DRC. Fortunately, the relational algebra, a second formal query language that shares the logical foundations of the DRC, aligns closely with SQL, perhaps the most widely used commercial data query language.

In this Chapter, we operationalize our model by extending the relational algebra to support attribution. We begin by sketching our intuition behind an algebra for attribution. Next, we provide some basic definitions from which we build the extended algebra. After presenting our attribution algebra, we consider some of the extended algebra's properties. We first show that the attribution algebra is closed. We then show that the extended algebra reduces to the standard relational algebra and is a consistent extension of the standard algebra (both properties are elaborated upon below). Finally, we prove that for algebraic expressions without nested negations, the attribution algebra supports the formal model. That is to say that for any algebraic query expression without nested negations, the extended algebra produces the relation-level source granules for attribute-value pairs in the result relation as defined by the formal model.

5.1 Algebra for attribution

In our extended algebra, metadata to calculate source, comprehensive, and relevant attribution is associated with attribute-value pairs of the relational data model. We propagate the attribution metadata in an eager fashion that updates source, relevant, and comprehensive attribution with each successive query operation.

In Chapter 2 on Related Work, we noted that eager approaches continuously maintain attribution values. While the overhead is higher, response to an attribution request is correspondingly faster. Purely lazy approaches, by contrast, wait until a request for attribution is posed. Depending upon the motivation, different applications might prefer one approach to the other. Because intellectual property provisions, as a matter of policy, apply uniformly, eager approaches may make the most sense. For data sets that are of generally high quality, a lazy approach for tracing anomalous values might be more appropriate.

For simplicity, we leverage the granularity intuition from Chapter 4. Associating attribution with each attribute-value pair corresponds to value-level result granules. Value-level result granularity preserves the observation that different attribute-values in the same tuple may draw from different sources and be subject to different constraints (source and relevant attribution). Conversely, rather than maintaining substitutions and query expressions, we propagate only relation names and query expressions. Relation-level source granularity certainly does not correspond to all of the different intuitions, but it both limits the amount of metadata maintained and propagated while satisfying the needs for specific attribution motivations. As argued earlier in our discussion of granularity, some issues such as remuneration or intellectual property are addressable by coarse-grained source granules.

5.2 Basic definitions

To present the extended algebra, we begin with a few basic definitions both as a brief review and as an introduction to the notation used throughout the remainder of this Chapter.

Let $\mathbf{D} = D_1 \cup D_2 \cup \ldots \cup D_n$ be the set of disjoint domains over which all relations are defined. A *scheme* is a pair (J, D) where J is an index (a set of integers) from 1 to $max(J)$ and D is a function that maps every element in the index to a domain in $\mathbf{D}$ $(D : J \rightarrow \mathbf{D})$. Note that in practice, this is no different than traditional attribute-value naming and is done here for notational convenience (Ullman 1988). A *relation* is then defined over a scheme as a finite subset of the Cartesian product of the domains in the scheme. Each element t of a relation R defined on scheme (J,D), written $t \in R$, is a tuple of scalars where for $j \in 1\ldots max(J)$, $t[j] \in D[j]$.

The *relational algebra* is then defined in terms of two unary and three binary operators that take one (or two in the case of binary operators) relations as arguments and returns a single relation. Domains in $\mathbf{D}$ are considered θ comparable meaning that we can evaluate the binary, Boolean operators $\{<, \leq, =, \geq, >\}$ for values in each domain.

Formal definitions of the unary and binary operators are given below. Here we offer more colloquial intuitions. *Select* (σ) is a unary operator that takes a relation R and a θ-condition. The resulting relation S is a subset of R containing all tuples of R that satisfy the θ-condition. *Project* (π) is a unary operator that takes a relation R on scheme (J,D) and a set of indexes $K \subseteq J$ specifying a subset of the domains in R. The resulting relation S contains unique tuples of R as defined by the projected domains (only values in domains $D[k]$).

Natural Join $(\bowtie)$ is a binary operator that concatenates tuples from each input relation R and S to create a single result tuple. For specified attribute domains that appear in both relations (e.g. as in the case of a foreign key), the duplicate occurrence is eliminated. Result tuples are those formed by R and S provided that the tuple from R and the tuple from S agree in the value(s) of all specified duplicate domains. *Union* (*union*) takes two relations R and S, defined on the same schema, and returns a relation containing all tuples in R and S. *Difference* (–) takes two relations defined on the same schema and returns those tuples that appear only in R.

Finally, throughout the remainder of this Chapter we refer to the *source* of a tuple or the source of the specific instance of a value (i.e. the unique tuple in which the referenced instance of a domain value appears) as a scalar representing the relation in which the tuple appears. A *source* is a relation name.

5.3 Extended algebra

5.3.1 Extended relation

We continue to define the set of all domains **D** and a relational scheme (J,D) as before. In the standard relation, each relation element is a tuple of scalars drawn from the corresponding domains. In an *extended relation*, however, every scalar is associated with two sets of sources and extended tuples are associated with an additional set of sources.

Definition 5.1 Extended relation (R')
An *extended relation* R' over scheme (J,D) is a finite subset of the Cartesian product of cells written $E_1 \times \ldots \times E_{max(J)} \times 2^S$. □

Definition 5.2 Extended tuple (t')
An element $t' \in R'$ is an *extended tuple* of R'. An extended tuple is a tuple of *cells* paired with a set of sources that returns the comprehensive attribution for every cell in the tuple. The j^{th} element of t' is the cell denoted by $t'[j]$ and the set of sources comprising the comprehensive attribution for the tuple is referenced as t_C'. □

Definition 5.3 Cell (E_j)
A cell is defined with respect to an extended relation R' on a schema (J,D). A cell is a triple composed of a scalar drawn from an attribute domain and sets of sources corresponding to the source attribution and relevant attribution for the scalar. For a scheme (J,D) and $j \in J$, we call E_j the Cartesian product $D[j] \times 2^S \times 2^S$. We reference these elements as $t_V[j]$, $t_S[j]$, and $t_I[j]$. □

Two or more tuples with identical values but different source sets are said to be weak duplicates. Such tuples are also referred to in the literature on extended algebras as value-equivalent tuples (Dey, Barron, and Storey 1996; Dey and Sarkar 1996).

Definition 5.4 Weak duplicate
Given two extended tuples t_1 and t_2 in extended relation R defined over the scheme (J,D), we say that t_1 and t_2 are weak duplicates if and only if $\forall j \in J,\ t_{1\,V}[j] = t_{2V}[j]$. □

5.3.2 Operations on extended relations

We now define a number of operations on extended relations from which we will construct our attribution algebra. From these operations we will define our attribution algebra for extended relations.

Definition 5.5 (δ) Weak duplicate elimination
Given an extended relation R' defined over the scheme (J,D), the removal of weak duplicates in R' is a relation over the scheme (J,D):
$S' = \delta(R') = \{s | \exists r \in R' \textit{ and } s_V[k] = r_V[k],\ s_S[k] = \bigcup_{\forall r \in dup(r)} r_S[k],\ s_I[k] = \bigcup_{\forall r \in dup(r)} r_I[k], \textit{ and } s_C = \bigcup_{\forall r \in dup(r)} r_C \}$. □

Weak duplicate elimination is very much like the coalesce function introduced by Snodgrass (Snodgress 1987 cited in: Bohlen, Snodgrass, and Soo 1996; Dey, Barron, and Storey 1996) to manage value equivalent tuples. Unlike much work in temporal databases, our (δ) is not an algebraic operator that users may use to manage overlapping temporal ranges.[28] Rather, we follow Wang and Madnick (1990) and Dey (1996), where weak duplicate elimination is incorporated into the extension of each algebraic operator's definition (see below) to preserve the relational set semantics, which does not allow weak duplicates.

The reader will note that a similar problem emerges with multiple relations involving the same attribute as in the case of a natural join on a foreign key or attributes used in a θ comparison as in select (σ). Because of the distinction noted previously in Chapter 4 between natural join on the same attribute domain and θ-comparable attribute domains, we provide for *attribute coalesce*.

Definition 5.6 (κ) Attribute coalesce
Given an extended relation R_1 over the scheme (J,D), a set $K \subseteq J$, the coalesce of R_1 for the attributes in K is the relation $R_2 = \kappa(R_1,L)$ over the domains in (J,D) such that, where $eq(t)$ is the application of the Boolean function verifying equality for all parameters on the values $t_v[k]$ of tuple t, $\forall k \in K$:

$R_2 = \kappa(R_1,K) = \{t_2 \mid \exists t_1 \in R_1$ such that $eq(t_1)$ and $\forall j \in J - K,\ t_2[j] = t_1[j]$ and $\forall j \in K$, $t_{2V}[j] = t_{1V}[j]$ and $t_{2S}[j] = \bigcup_{\forall k \in K} t_{2S}[k],\ t_{2I}[j] = \bigcup_{\forall k \in K} t_{2I}[k],\ t_{2C} = t_{1C}\}$ □

Definition 5.7 (σ^+) Select$^+$
Given an extended relation R_1 over the scheme (J,D), a set $K \subseteq J$, and a Boolean function θ over the domain $D(k_1) \times \ldots \times D(k_K)$ the selection of R_1 on the condition θ for the attributes $k \in K$ is $R_2 = \sigma(R_1,\theta,L)$ over the domain (J,D) such that, where $\theta(t)$ is the application of the

[28] (Dey, Barron, and Storey 1996) provides a nice review of different coalesce operators in the literature to manage time stamps

Boolean function θ on the values $t_V[k]$ of tuple t, we define a function *Relevant*(Y) that returns the set of variables relevant to the set of domain variables Y and set $X = Relevant(K)$.

$R_2 = \sigma(R_1,\theta,L) = \{t_2 | \exists t_1 \in R_1$ such that $\theta(t_1)$ and $\forall j \in J$, $t_{2V}[j] = t_{1V}[j]$, $t_{2S}[j] = t_{1S}[j]$, $t_{2C} = t_{1C}$ and if $j \in Relevant(K)$ then $t_{2I}[j] = t_{1I}[j] \cup \bigcup_{k \in K} t_S[k] \cup \bigcup_{k \in K} t_I[k]$ else $t_{2I}[j] = t_{1I}[j]\}$ □

Relevant is recursively defined to identify all sources that are mutually dependent through θ-comparisons. The set I updates which values in the tuple of an extended select are bound by evaluating the θ-condition. In this way, we make explicit the observation that the θ-condition is *relevant* to specific values in the corresponding tuple of the result relation.[29]

Definition 5.8 (π^+) Project$^+$

Given an extended relation R_1 over the scheme (J_1,D_1), an index J_2, and a function p from J_2 to J_1, the projection of R_1 w.r.t. p is $R_2 = \pi(R_1)$ over the scheme (J_2,D_2) such that:
$\forall j \in J_2$, $D_2(j) = D_1(p(j))$, and
$R_2 = \pi(R_1) = \delta(\{t_2 | \exists\, t_1 \in R_1$ and $t_{2C} = t_{1C}$ and $\forall j \in J_2$, $t_2[j] = t_1[p(j)]\})$. □

Definition 5.9 ($\times^+$) Cartesian Product$^+$

Given two extended relations R_1, defined over the scheme (J_1,D_1), and R_2, defined over the scheme (J_2,D_2), the Cartesian product$^+$ of R_1 and R_2 is a relation $R_3 = R_1 \times R_2$ over the scheme (J_3,D_3) such that, for $M_1 = max(J_1)$ and $M_2 = max(J_2)$:
J_3 is an index ranging from 1 to $M_1 + M_2$, and
$\forall j \in J_3$, if $j \leq M_1$ then $D_3(j) = D_1(j)$, else $D_3(j) = D_2(j - M_1)$, and
$R_3 = R_1 \times R_2 = \{t_3 | \exists t_1 \in R_1$ and $\exists t_2 \in R_2$ and $t_{3C} = t_{1C} \cup t_{2C}$ and $\forall j \in J_3$, if $j \leq M_1$ then $t_3[j] = t_1[j]$ else $t_3[j] = t_2[j - M_1]\}$ □

Definition 5.10 ($-^+$) Difference^{+}[30]

Given two extended relations R and S defined over the scheme (J,D), the difference of R and S is a relation $T = R - S$ over the scheme (J,D) such that $T = R - S = \{t | \nexists s \in S$ such that $\forall j$, $t_V[j] = s_V[j]$ and $\exists r \in R$ such that $\forall j \in J$, $t_V[j] = r_V[j]$, $t_S[j] = r_S[j]$, $t_I[j] = r_I[j] \cup \bigcup_{\forall s \in S} s_C$ and $t_C = r_C \cup \bigcup_{\forall s \in S} s_C$. □

[29] We introduced the function *Relevant* rather than explicitly defining the term because of our difficulty in either explicitly defining the term or in characterizing how tightly our syntactic rule bound the formal definition of relevance. We present the following as one bound on relevance: *relevant*($t_S[k]$) is initialized to $\bigcup_{k \in K} t_S[k]$ and recursively defined as *relevant*($t_S[k]$) $\cup\ t_S[j]$ where $t_S[j] \cap$ *relevant*($t_S[k]$) is not empty.

[30] As will be discussed in greater detail below, the treatment of algebraic difference differs from our management of negation in the formal model of Chapter 4. However, for algebraic expressions without nested negations, we will see that the algebra and the formal model agree.

The set of sources t_l captures our intuition about negation. To verify that some instance of a value (e.g. the value in a specific extended tuple) does *not* exist in some extended relation S', we must compare the value-instance to every valid substitution in S'.

5.3.3 Extended relational operators

Building from the operators defined on extended relations, we can now define the attribution algebra as an extension of the standard relational algebraic operators. The attribution for an expression is then defined inductively from the extended definitions of the operators.

Definition 5.11 (σ') Extended select
Given an extended relation R', $\sigma'(R',\theta,L) = \sigma^{+}(R',\theta,L)$ □

The extended select is simply the select defined on extended relations.

Definition 5.12 (π') Extended project
Given an extended relation R', $\pi'(R') = \delta(\pi^{+})$ □

The extended project is a projection followed by a weak duplicate elimination in order to account both for duplicates among extended tuples and duplicates among value equivalent tuples.

Definition 5.13 ($\bowtie'$) Extended natural join
Given extended relations R' and S' defined on schemas (J_1,D_1) and (J_2,D_2) respectively with a function p that maps $H \subseteq J_1$ to J_2 such that $D_1(h) = D_2(p(h))$,
$R' \bowtie' S' = \kappa(\sigma(R' \times^{+} S', \theta(=), \forall H), \forall H)$ □

The extended natural join is a Cartesian product on extended relations followed by a selection on equality for all attribute domains used (named) identically as indicated by the function p. Finally, we coalesce on all attribute domains used (named) identically. The reader may observe that the effect of an extended Cartesian product ($\times'$) is achieved by taking the extended natural join where H is empty. Likewise, extended Intersection ($\cap'$) is simulated by taking extended natural join on two relations R' and S' defined for the same schema (J,D).

Definition 5.14 ($\cup'$) Extended union
Given extended relations R' and S' defined on the same schema (J,D), the extended union R' $\cup' S' = \delta(R' \cup S')$ where $\cup$ is the standard set union operator. □

Extended union is simply the standard set union operator that uses weak duplicate elimination to manage value equivalent tuples with different sets of sources.

EXTENDED ALGEBRA

Definition 5.15 (–') Extended difference

Given extended relations R' and S' defined on the same schema (J,D), the extended difference $R' -' S' = R' -^{+} S'$ □

We can now define attribution in the context of our extended relational algebraic operators. As we define attribution, we informally relate our algebraic definitions to the formal model of Chapter 4. A formal proof of the relationship between the algebraic definition and the formal model is provided later.

Definition 5.16 Comprehensive attribution

The comprehensive attribution for a scalar $t_V[j]$ in the result of an extended relational algebraic expression E having schema (J,D) is defined as the set t_C. □

t_C is in fact the comprehensive attribution for the entire tuple reflecting the observation from the formal model that when considering relation-level source granules, the comprehensive substitutions that make any value of tuple t in the expression true are the same for every other value in tuple t. Moreover, managing the *difference* operator is actually captured in t_C by construction. This explains Definition 5.15 that updates t_C with the comprehensive attribution for every tuple of the negated relation when evaluating the difference of extended relations R and S.

Definition 5.17 Source attribution

The source attribution for a scalar $t_V[j]$ in the result of an extended relational algebraic expression E having schema (J,D) is defined as the set $t_S[j]$. □

The attribution algebra continuously updates the source attribution for each scalar value in an extended relation by managing the set $t_S[j]$. Note that the source attribution for a value in a tuple is not updated by the extended project or extended union except in the case of weak duplicates. In these instances, weak duplicates represent multiple occurrences of an instance in the same relation (project) or distinct derivations for the same instance (union) as discussed in the formal model. Likewise, source attribution is not updated in the case of natural join except for those values that are drawn from the same (named) attribute domain (i.e. coalesced). In the formal model, we identified this as multiple occurrences of the same variable in different conjuncts representing relational predicates. Note also how the set $t_S[j]$ is not altered in the definition of extended set difference, corresponding to our intuition that a negated sub-query is never a source for a value in the result of the difference.

Definition 5.18 Relevant attribution

The relevant attribution for a scalar $t_V[j]$ in the result of an extended relational algebraic expression E having schema (J,D) is defined as the set $t_S[j] \cup t_I[j]$. □

Notice that the relevant attribution is defined in terms of two sets of sources, $t_S[j]$ and $t_I[j]$. The set $t_I[j]$ is not updated for extended project and extended union except in the case of weak duplicates. Because weak duplicates represent distinct derivations for a given instance of a value in the result, we legitimately include the relevant attribution for each weak duplicate. We see that $t_I[j]$ is always updated when evaluating the extended difference but only selectively updated when evaluating θ-conditions.

For extended difference, $t_I[j]$ is updated with the *comprehensive attribution* of *every* tuple in the negated relation. Comprehensively attributing every tuple corresponds to our intuition from the formal model about evaluating the truth of a negated sub-formula. We see that relevant attribution includes $t_S[j]$ corresponding to the idea that the source of a value is certainly relevant.

In the selection operation, we update the relevant attribution for every value in a tuple with the relevant attribution of the selection variables. Intuitively, a selection condition restricts a subset of (possibly all) values in the result tuple hence the introduction of the *relevant* function which relations are linked through θ-comparison. Recall also the implicit selection-on-equality in the natural join. Note that in the special case of natural join where there are no shared variables (i.e. no implicit selection), the relevant attribution for values in the result are drawn exclusively from the corresponding constituent tuple of the Cartesian product. This corresponds to our intuition from the formal model that restricting the tuples in one argument of a Cartesian product is not a restriction on the second argument.

5.4 Properties of the algebra

Having presented our attribution algebra, we now consider properties of the extended algebra. We demonstrate first that the algebra is closed. Then, following the literature on extended algebras for temporal databases (Dey, Barron, and Storey 1996), we establish that the attribution algebra both reduces to and is a consistent extension of the standard relational algebra. Finally, we show that, for a limited set of extended algebraic query expressions, the attribution returned by the algebra corresponds to the relation-level source granules defined by the formal model.

5.4.1 Closure of the extended algebra

The intuition behind closure is that an extended algebraic operation, when applied to an extended relation(s), returns an extended relation. Maier (1983) identifies three requirements:

1. the values in each cell of the extended relation all come from the correct domains
2. there are no (weak) duplicates in an extended relation
3. the relations must be finite

EXTENDED ALGEBRA

Lemma 5.1 The values in each cell of the output from an extended operation on extended relation(s) all come from the correct domains.

Case ($\pi'(R')$ where R is defined on schema (J,D)): We know by definition that $\pi'(R')$ is defined on a schema (K,D) where $K \subseteq J$ and that for every $s' \in \pi'(R')$ $\exists t' \in R'$ such that $\forall k = p(j) \in K$, $s'_V[p(j)] = t'_V[j]$ so all values come from valid domains. In cases where there are no weak duplicates, then $s' \in \pi'(R')$ is value equivalent to exactly one tuple $t' \in R'$. In this case, $s'_C = t'_C$ and $\forall k = p(j) \in K$, $s'_S[p(j)] = t'_S[j]$ and $s'_I[p(j)] = t'_I[j]$ so the sets of scalars all come from the appropriate domains. If there are weak duplicates among the $t'_V[j]$ for all $j \in K$, then the sets s'_C, s'_S, and s'_I are simply the union of the constituent weak duplicates and the union of valid scalar sets is surely still in 2^S.

Case ($\sigma'(R')$): We assume that R' is an extended relation. Therefore, we know that $t' \in \sigma'(R') \rightarrow t' \in R'$ so if R' is an extended relation, then $\sigma'(R')$ must also.

Case ($(R' \cup' S')$ where R and S are union compatible in the standard sense on schema (J,D)): We know that an extended tuple $t' \in R' \cup' S'$ must come from R', from S', or from both. Consider first the case where t' comes from only one. Then we know for such a tuple t', $\exists r' \in R'$ or $\exists s' \in S'$ such that $t' = r'$ or $t' = s'$ and all values come from appropriate domains. In the case that t' comes from both, then we know, as with weak duplicates in project, that $t'_C = r'_C \cup s'_C$ and $\forall j \in J$, $t'_V[j] = r'_V[j] = s'_V[j]$, $t'_S[j] = r'_S[j] \cup s'_S[j]$ and $t'_I[j] = r'_I[j] \cup s'_I[j]$.

Case ($R' \bowtie' S'$): Recall from the definition that this is a Cartesian product followed by a selection and a coalesce on the common attributes $K \subseteq J$. Certainly the Cartesian product of extended relations is an extended relations because it is merely the $r' \circ s'$ for every $r' \in R$ and $s' \in S'$. Likewise, the select also returns an extended relation (see above). Consider, then, the Coalesce. $\forall t \in R' \bowtie' S'$, t_C is unchanged from the Cartesian product and select. For indexes $j \notin K$ we know that $t[j]$ is unchanged from the Cartesian product and select. For index in K, we know that $t_V[k]$ is unchanged and that $t_S[k]$ and $t_I[k]$ is the union of all values in K where each t_S and t_I is from the correct domains. Hence the union must still be in 2^S.

Case ($(R' - S')$ where R and S are union compatible in the standard sense on schema (J,D). For $t' \in (R' -' S')$, $\exists r' \in R'$ such that $t'_C = r'_C \cup \bigcup_{\forall s' \in S} s'_C$ so surely t'_C is from the correct domain. Moreover, $\forall j \in J$, $t'_V[j] = r'_V[j]$ and $t'_S[j] = r'_S[j]$. By construction, $t'_I[j]$ is the union of valid source sets, hence we conclude that the values in each cell of the output from an extended operation on extended relation(s) all come from the correct domains. □

Lemma 5.2 There are no (weak) duplicates in the output of an extended operation on extended relation(s).

First, we know that extended relations are defined as sets so that there are no duplicate extended tuples in an extended relation. A different question is whether the extended operators can produce weak duplicates. We know from their definitions directly that extended select, extended join, and extended difference cannot produce weak duplicates assuming that the initial input relation(s) are valid extended relations (i.e. with no (weak) duplicates). The remaining operators, extended union and extended project both are defined as explicitly calling weak duplicate elimination. Hence, we are assured that there are no (weak) duplicates in the output of an extended operation on extended relation(s). □

Lemma 5.3 The result of an extended operation on extended relation(s) is finite.
Case ($\pi'(R')$ where R' is defined on schema (J,D)): We know that $|\pi'(R')| \leq |R'|$ because each extended tuple of $\pi'(R')$ is a tuple of R' on (K,D) where $K \subseteq J$. At most, every tuple of $\pi'(R')$ is distinct, reduced by weak duplicate elimination. Therefore if R' is finite, $\pi'(R')$ must also be finite.

Case ($\sigma'(R')$): By definition, $\sigma'(R') \subseteq R'$ therefore $|\sigma'(R')| \leq |R'|$.

Case (($R' \cup' S'$) where R' and S' are union compatible in the standard sense): It must be the case that $|R' \cup' S'| \leq |R'| + |S'|$. If R' and S' are both finite, then so is $R' \cup' S'$. Note as in the case of extended project, weak duplicates will reduce the cardinality of $R' \cup' S'$.

Case ($R' \bowtie' S'$): This is a Cartesian product followed by a select and a coalesce. As observed above, an extended select either leaves the cardinality of the input relation unchanged or reduces it. Coalesce merely collapses duplicate attribute (domains); the output of a coalesce has the same cardinality as the input. Thus, we conclude $|R' \bowtie' S'| \leq |R'| \times |S'|$.

Case (($R' -' S'$) where R' and S' are union compatible in the standard sense): Thus $R' -' S' \subseteq R'$ so $|R' -' S'| \leq |R'|$. □

Theorem 5.1 The attribution algebra is closed.
From Lemmas 5.1-3, we conclude that Theorem 5.1 holds. □

5.4.2 Relationship between the standard algebra and extended algebra

Having verified that we can compose operators, we next verify that the extended algebra is both a consistent extension of and reduces to the standard algebra. When we say that the extended algebra reduces to the standard algebra, we are saying that the extended algebra preserves the relational semantics. In other words, from the perspective of the scalar values drawn from attribute domains, the extended operators treat an extended relation on schema (J,D) as the standard relation would treat the corresponding standard relation on the same schema and for the same attribute-value substitutions. Following Dey (1996; 1996), we first define a helper function *Reduce*. The purpose of *Reduce* is to take an extended relation and map it to the equivalent relation without the attribution extension. We then show that the extended algebra reduces to the standard algebra through an equivalence proof. The equivalence proof is illustrated in Figure 5.1.

Definition 5.19 Reduce

Given an extended relation R' on a schema (J,D), $reduce(R') = \{t_2 \mid \exists t_1 \in R'$ and $\forall j \in J,\ t_2[j] = t_{1V}[j]\}$ also on scheme on (J,D).

$$\begin{tikzcd}
 & & S'\,\theta'\,R' \\
S' \arrow[r, dashed] \arrow[d, "Reduce(S')"] & R' \arrow[r] \arrow[d, "Reduce(R')"] & \theta'(R') \arrow[d] \\
S \arrow[r, dashed] & R \arrow[r] & \theta(R) = Reduce(\theta'(R')) \\
 & & S\,\theta\,R = Reduce(S'\,\theta'\,R')
\end{tikzcd}$$

Figure 5.1 Reduction

Theorem 5.2 The extended algebra reduces to the standard algebra

To prove the theorem, we need simply show that the reduction holds for every unary and binary operator of the extended algebra. In each case, we need to show both directions. The reduction of a tuple $t' \in$ extended operator is in a standard operator applied to reduced inputs and vice versa.

Case ($\pi'(R')$ where R' on schema (J_1,D_1)): By definition of *Reduce* we know that R is also defined on (J_1,D_1) and by definition of extended project, we know that $\pi'(R')$ is defined on a function p and produces a schema (J_2,D_2). Note that $\pi(R)$ is defined similarly for R on (J_1,D_1) and the same p. Assume that $R = Reduce(R')$. Pick some $t \in \pi(R)$. Then by definition of π, $\exists\ t_1 \in R$ s.t. $\forall j \in J_2,\ t_1[p(j)] = t_2[j]$. Because t_2 is a set, we know that there may be more than one such t_1, but there is certainly at least one. From the definition of *Reduce*, we know that for t_1 on (J_1,D_1), $\exists\ t_1' \in R'$ such that $\forall j \in J_1,\ t_1'{}_V[j] = t_1[j]$. But then $\pi'(R')$ must give t_2' on J_2,D_2 where $\forall j \in J_2,\ t_2'[j] = t_1'[p(j)]$ by definition of π'. And because $t_1 = Reduce(t_1')$, certainly $Reduce(t_1'[p(j)]) = t_1[p(j)]$. This tells us that $Reduce(t_2') = t_2$ so $t_2 \in Reduce(\pi'(R'))$. Likewise, pick some $t_2' \in \pi'(R')$ where we know $Reduce(t_2')$ gives t_2 on (J_2,D_2) when $\forall j,\ t_2[j] = t_2'{}_V[j]$. By definition of π' we know $\exists t_1' \in R'$ such that $\forall j \in J_2,\ t_1'[p(j)] = t_2'[j]$ where there may be more than one such t_1 on (J_1,D_1). But $Reduce(t_1') = t_1$ on $(J_1,D_1) \in R$ where $\forall j \in J,\ t_1'{}_V[j] = [j]$. This means that $t_1[p(j)] = t_1'{}_V[p(j)]$ or that $t_2 = t2'_V\ \forall j \in J_2$.

Case ($\sigma'(R')$ where R' is defined on schema (J,D)): Pick $t \in \sigma(R)$ and assume $\neg\exists t' \in \sigma'(R')$ for which $t = Reduce(t')$. We know that $R = \{t \mid \exists t' \in R'$ and $\forall j\ t[j] = t'_V[j]\}$ so for every t, there must be some t'. But t satisfies (θ,L) which means $\forall k,\ t_V'[k]$ also satisfies θ, a contradiction. Now pick $t' \in \sigma'(R')$ where $t = \mathit{Reduce}(t')$. Then assume $\neg\exists t \in \sigma(R)$. But if $t' \in \sigma'(R')$ then t'_V satisfies (θ,L). But $\forall j \in J$, $t'_V[j] = t[j]$ by definition of reduce so t must also satisfy (θ,L) which means that $t \in \sigma(R)$, a contradiction.

Case $((R' \cup' S')$ where R' and S' are union compatible in the standard sense): Pick $t' \in R' \cup' S'$. If we $\mathit{Reduce}(t')$ we get t where $\forall j \in J$, $t[j] = t'_V[j]$. But by definition, we know $t' \in R'$, $t' \in S'$, or both. If $t' \in R'$ then $\mathit{Reduce}(t') = t \in R$ by definition which means $t \in R \cup S$. Likewise for $t' \in S'$ so certainly for both. Now pick $t \in R \cup S$. Then $t \in R$, $t \in S$, or both. When $t \in R$, we know $\exists t' \in R'$ *s.t.* $\forall j \in J$, $t'_V[j] = t[j]$ meaning that $t = \mathit{Reduce}(t')$. But if $t' \in R'$ then $t' \in R' \cup' S'$ and the same for $t \in S$ and again certainly for both.

Case $(R' \bowtie' S')$: Pick $t \in \mathit{Reduce}(R' \bowtie' S')$. Then t corresponds to $t' \in R' \bowtie' S'$ where $\forall j\; t[j] = t'_V[j]$. (J,D) is the schema for $R' \bowtie' S'$. Then $\forall j$, $t'_V[j]$ is from R' or from S' or from both (if j is in the k's of overlapping domains from which the selection on the Cartesian product is made). But for R', $t'_V[j] = t[j] \in R$. Likewise for S' and S. We note that for $t'_V[k]$, $t[k]$ holds in R and S. Certainly $t \in R \bowtie S$. Now pick $t \in \mathit{Reduce}(R') \bowtie \mathit{Reduce}(S')$. Then $\forall j$, $t[j]$ from $\mathit{Reduce}(R')$, $\mathit{Reduce}(S')$ or both in the event that j is in the k's). From the definition of *Reduce*, we see that $t'_V[j] = t[j]$ in R' and similarly for S'. Finally, for the k's, we see that $t'_V[k] \in R' = t'_V[k] \in S'$. Hence we conclude that $\mathit{Reduce}(R' \bowtie' S') = \mathit{Reduce}(R') \bowtie \mathit{Reduce}(S')$.

Case $((R' -' S')$ where R' and S' are union compatible in the standard sense. Pick $t \in \mathit{Reduce}(R' -' S')$. Then t corresponds to $t' \in R' -' S'$ where $\forall j\; t[j] = t'_V[j]$. Then $\forall j$, $t'_V[j] \in R'$ and $\notin S'$. Surely $\mathit{Reduce}(t') = t \in R$. And if $t' \notin S'$ then $\mathit{Reduce}(t') = t \notin S$. So, we know that $t \in \mathit{Reduce}(R' -' S')$ appears in $\mathit{Reduce}(R') - \mathit{Reduce}(S')$. Now pick $t \in \mathit{Reduce}(R') - \mathit{Reduce}(S')$. Then $\forall j$, $t_V[j] \in \mathit{Reduce}(R')$ and $\notin \mathit{Reduce}(S')$. Then $\exists t' \in R'$ such that $\forall j$, $t'_V[j] = t[j] \in R$ and $t' \notin S'$. Thus, we see that $\mathit{Reduce}(R' -' S') = \mathit{Reduce}(R') - \mathit{Reduce}(S')$.

Therefore, we may conclude that for unary operators, $t \in \mathit{Reduce}(op'(R'))$ iff $t \in op(\mathit{Reduce}(R'))$ and for binary operators, $t \in \mathit{Reduce}(R\ op\ S)$ iff $t \in \mathit{Reduce}(R)\ op'\ \mathit{Reduce}(S)$.
□

Having verified that the extended algebra reduces to the standard algebra, we consider the inverse and ask whether the extended algebra is a consistent extension of the standard algebra. In other words, we are asking whether the attribution algebra has the property that every relational algebra expression has a counterpart in the extended algebra. Again following Dey (1996; 1996), we first define a helper function *Extend*. Extend takes an algebraic expression as a single argument and extends the corresponding relation by applying the formal model to the DRC equivalent assuming a database of relations in the original argument. Because there may be more than one valid extended form for a relation (e.g. depending upon the database against which an expression is evaluated), we again turn to an equivalence proof. To demonstrate that the algebra is a consistent extension, we want to show that extending the extended relational operation on the extended relational inputs. This intuition is depicted in Figure 5.2.

Definition 5.20 Extend
Given an algebraic expression E that returns a relation R on schema (J,D), *Extend* transforms E into its DRC equivalent F[31] having formula f to construct the extended relation R'. Let database d be comprised of the relations in the expression E and *granularity* (A) take an attribution and return the relation names corresponding to the substitutions. Then $Extend(R)$ $= \{t_2 \mid \exists t_1 \in R$ where $t_{2C} = granularity(comprehensive\text{-}attribution(t_1, F, d)$ and $\forall j \in J$,

$t_{2V}[j] = t_1[j]$
$t_{2S}[j] = granularity(source\text{-}attribution(t_1[j], F, d)$
$t_{2I}[j] = granularity(relevant\text{-}attribution((t_1[j], F, d)\}$ □

$$\begin{tikzcd}
S \arrow[r, dashed] \arrow[d, "Extend(S)" description] & R \arrow[r] \arrow[d, "Extend(R)" description] & \begin{array}{c} S\,\theta\,R \\ \theta(R) \end{array} \arrow[d] \\
S' \arrow[r, dashed] & R' \arrow[r] & \begin{array}{c} \theta'(R)' = Extend(\theta(R)) \\ S'\,\theta'\,R' = Extend(S\,\theta R) \end{array}
\end{tikzcd}$$

Figure 5.2 Extension

Theorem 5.3 The extended algebra is a consistent extension of the standard algebra
As with reduction, we show that each extended operation is a consistent extension of its standard analog. Let E be an abbreviation for the function *Extend.*

Case $(\pi'(R'))$: Pick $t' \in \pi'(E(R)$: By definition, $t'_C = \{R\}$. $\forall j\ t'_V[j] = t[p(j)]$ which is just $\pi(R)$. $t'_S[j] = \{R\}$. $t'_I[j] = \varnothing$. Then certainly $t' \in E(\pi(R))$. Now pick $t' \in E(\pi(R))$. Then $\exists t \in \pi(R)$ for which $t'_V[j] = t[j]$. If we extend t into some t', we know that $\forall j$, $t'_S[j] = \{R\}$, $t'_I[j] = \varnothing$, and $t'_C = \{R\}$. Of course this is just $\pi'(E(R))$.

Case $(\sigma'(R')$ for the selection condition θ,K where R' is defined on schema $(J,D))$: Recognizing that the selection condition θ,K is the same for both σ and σ', we define the set X $= Relevant(K)$ for both the standard and the extended select. Pick $t' \in \sigma'(E(R))$. By definition, θ,K is true for all t'. Furthermore, we know $\exists t \in R$, $\forall j\ t'_V[j] = t[j]$; $t'_S[j] = \{R\}$; $t'_C =$

[31] See (Ullman 1988)

$\{R\}$ and $\forall x \in X$, $t'_I[x] = \bigcup_{\forall x} t'_I[x] \cup t'_S[x]$. $\forall j \notin X$, $t'_I[j] = \varnothing$.[32] But the set of all such t is simply $\sigma(R)$ which extended is the set of all t'. This is just $E(\sigma(R))$. Similarly, we can pick $t' \in E(\sigma(R))$ which is just the extension of $t \in \sigma(R)$. Then we know that $E(\sigma(R))$ gives t' such that $t'_C = \{R\}$ and $\forall j\ t'_V[j] = t[j]$; $t'_S[j] = \{R\}$; $t'_C = \{R\}$ and for X (in this case $X = J$), $\forall x \in X$, $t'_I[x] = \bigcup_{\forall x} t'_I[x] \cup t'_S[x]$. $\forall j \notin X$, $t'_I[j] = \varnothing$. But this tuple is certainly in $E(R)$ because $\sigma(R) \subseteq R$ and we know that t satisfies θ, K as does t'. Therefore we know $t' \in \sigma'(E(R))$.

Case $((R' \cup' S')$ where R' and S' are union compatible in the standard sense): If $t' \in E(R) \cup' E(S)$ then $t' \in E(R)$, $t' \in E(S)$, or both. If $t' \in E(R)$ then $\exists t \in R$, $\forall j\ t'_V[j] = t[j]$; $t'_S[j] = \{R\}$; $t'_I[j] = \varnothing$ and by definition, $t'_C = \{R,S\}$. Certainly if $t \in R$, $t \in R \cup S$. Moreover, because $t' \notin E(S)$ then we know there is no t'' in $E(S)$ for which $\forall j\ t''[j] = t'[j]$. Thus, we know that $t' \in E(R \cup S)$. The same holds for $t' \in E(S)$. Now suppose $t' \in E(R)$ and $E(S)$. Then $\exists t_1 \in R$ and $\exists t_2 \in S$. If we were to extend t_1 and t_2 we would find that for $t = t_1 \cup t_2$ when $\forall j\ t_1[j] = t_2[j]$, $t'_V[j] = t_1[j] = t_2[j]$; $t'_S[j] = t_1'_S[j] \cup t_2'_S[j] = \{R,S\}$; and $t'_I[j] = t_1'_I[j] \cup t_2'_I[j] = \{\}$. $t'_C = t_1'_C \cup t_2'_C = \{R,S\}$. But for $t_1 = t_2$ certainly $t \in R \cup S$ hence $t' \in E(R \cup S)$. Now, if $t' \in E(R \cup S)$ then we know that $\exists t \in (R \cup S)$ s.t. $t'_C = \{R,S\}$ and $\forall j\ t[j] = t'_V[j]$. If $t \in R$ then $t'_S[j] = \{R\}$. Likewise if $t \in S$. If $t \in R$ and $t \in S$ then we know $t'_S[j] = \{R, S\}$. But if $t \in R$ (and not S) then $\exists t_1' \in E(R)$ and there is no $t_2' \in E(S)$ so we know that for $t' = t_1' \cup' t_2'$, $t' \in E(R) \cup' E(S)$. We can say the same if $t \in S$ and not in R. If $t \in R$ and $t \in S$ then we know $\exists t_1' \in E(R)$ and $t_2' \in E(S)$ for which $t' = t_1' \cup' t_2'$. Then $\forall j\ t_1[j] = t_2[j]$, $t'_V[j] = t_1[j] = t_2[j]$; $t'_S[j] = t_1'_S[j] \cup t_2'_S[j] = \{R,S\}$; and $t'_I[j] = t_1'_I[j] \cup t_2'_I[j] = \{\}$. $t'_C = t_1'_C \cup t_2'_C = \{R,S\}$. Thus, we know $t' \in E(R) \cup' E(S)$.

Case $(R' \bowtie' S')$[33]: Let R be defined in J_1,D_1 and S be defined on J_2,D_2 with $n = max(J_1)$ and $m = max(J_2)$. The result of the natural join is a relation on scheme J,D where $J \leq n + m$. K is the set of selection attributes where $K \subseteq \{1 \ldots n + m\}$ and p is the projection function for J to $\{1 \ldots n + m\}$. First, assume $K = \varnothing$. Natural join then reduces to Cartesian Product. Pick $t' \in E(R) \bowtie' E(S)$. Then we know that $\forall j_{1..n}\ t'[j] = t_1'[j] \in E(R)$ and that $\forall j_{n+1..n+m}\ t'[j] = t_2'[j - n] \in E(S)$. Finally, $t'_C = \{R,S\}$. But then $t'_1 \in E(R)$ corresponds to $t_1 \in R$ and likewise for t_2' and $t_2 \in S$. Thus we see $t \in R \bowtie$ S and $t' \in E(R \bowtie S)$. If $t' \in E(R \bowtie S)$ then we know $\exists t \in R \bowtie S$ s.t. $\forall j_{1..n}\ t[j] = t_1[j] \in R$ and that $\forall j_{n+1..n+m}\ t[j] = t_2[j - n] \in S$. But we can extend t_1 to $t_1' \in E(R)$ and likewise for $t_2' \in E(S)$. We construct t' from t_1' and t_2' s.t. $t'_C = \{R,S\}$. Thus, we know $t' \in E(R) \bowtie' E(S)$. Now, we assume $K \neq \varnothing$. We then make use of Theorem 5.1 and the earlier cases for Cartesian product, selection, and then finally projection to verify that the

[32] In this instance, for $x \in X$, $t'_I[x]$ is just $\{R\}$. Otherwise, $t'_I[j] = \varnothing$. Note that in the more general case (as in the inductive case considered later in this Chapter), the Intermediate set for $t'_I[j]$ of $t' \in \sigma'(R')$ is by default the intermediate value for the corresponding $t'_I[j] \in R'$. As noted earlier, we introduced the function *Relevant* as a proxy for a syntactic rule.

[33] Recall that we define natural join as a Cartesian product followed by a selection on equality for attributes on the same domain, a coalsce, and then a projection of the duplicate columns. If there are no join attributes, then we simply have a Cartesian product. If the two schemas are the same, then we have an intersection.

property holds. In particular, we know that $\forall k \in K$, $t'_S[k] = \{R,S\}$ and that for each $k \in K$, we know that $x \in Relevant(K)$ as in the selection condition. In this instance, $t'_I[x] = \{R,S\}$. For $j \notin K$, $t'_S[j] = \{R\}$ or $\{S\}$ depending upon whether $p[j] \leq n$. Likewise for $j \notin Relevant(K)$, $t'_S[j] = t_1{}'_I[p(j)]$ or $t_2{}'_I[p(j) - n]$ depending upon whether $p[j] \leq n$.

Case $((R' -' S')$ where R' and S' are union compatible in the standard sense): If $t' \in E(R) -' E(S)$ then t' is an extended tuple $t' \in E(R)$ and $t' \notin E(S)$. This means that $\exists t$ s.t. $t \in R$, $t \notin S$, and $\forall j\ t'_V[j] = t[j]$; $t'_S[j] = \{R\}$; $t'_I[j] = \{S\}$; $t'_C = \{R,S\}$. But then $t \in (R - S)$ and it is easy to see that extending t we get $t' \in E(R - S)$. Now pick $t' \in E(R - S)$. Then $\exists t$ s.t. $t \in R$, $t \notin S$ and $t'_C = \{R,S\}$. $\forall j\ t'_V[j] = t[j]$; $t'_S[j] = \{R\}$; $t'_I[j] = \{S\}$. But if $t \in R$ and $t \notin S$, we can extend t to $t'' \in E(R)$ and we know that $t'' \notin E(S)$. It is then easy to see that $t'' = t' \in E(R) - E(S)$.

Therefore, we may conclude that for unary operators, $t' \in E(op(R))$ iff $t' \in op'(E(R))$ and for binary operators, $t' \in E(R\ op\ S)$ iff $t' \in E(R)\ op'\ E(S)$. □

5.4.3 Relationship between the extended algebra and the formal definition

Having related our attribution algebra to the standard relational algebra, we finally consider the relationship between the extended algebra and the formal model of Chapter 4. In particular, we want to know whether the extended algebra supports attribution as defined in the formal model.

From Theorem 5.2, we know that we can translate query expressions in the extended algebra into equivalent expressions in the standard algebra. From Ullman (1988), we know that we can translate algebraic query expressions into equivalent queries in the Domain Relational Calculus. Therefore, for any query expression in the extended algebra, using the DRC translation of Ullman (1988), we can evaluate whether the relations in the algebraic attribution correspond to the substitutions in the formal model for comprehensive, source, and relevant attribution. The comparison confirms that for algebraic query expressions without nested subtraction in the right hand side of a difference expression (the subtrahend), the algebraic attribution corresponds to the formal model.

We saw in Chapter 4 that because of its additivity property, attribution has complications when faced with *nested negations* (i.e. $x = \neg(\neg x)$). To account for this limitation, we first verify:

Lemma 5.4 Nested negations
Algebraic query expressions without nested subtraction in the right hand side of a difference expression correspond to Disjunctive Normal Form DRC expressions where negations are pushed down to literals without *nested negations* (e.g. canceling $\neg(\neg x)$). We establish this by induction on the number of operators in the algebraic expression.

In the base case of zero operators, the algebraic query expression is a single relation R on schema (J,D) or a constant relation. We know from Ullman (1988) that this is translated into an equivalent relational predicate $r(X_1,\ldots,X_{max(J)})$ or a corresponding expression for the constant relation $\{t_1,\ldots,t_n\}$ on (J,D) with formula $(X_1 = t_1D(1) \wedge\ldots\wedge X_{max(J)} = t_1D(max(J)))$ $\vee\ldots\vee$ $(t_nD(1) \wedge\ldots\wedge t_nD(max(J)))$ where there is a disjunct for each tuple t_i. Certainly in the base case there are no nested negations.

In the induction hypothesis, we assume that for a query with n operators, assuming no difference operators in the right-hand sub-tree of a difference operator, the resulting DRC translation in DNF with negations pushed down to literals will not nest negations. We want to verify that the same holds for a query expression with $n+1$ operators.

Case $(\pi(R))$: The DRC expression for the projection merely reassigns the set of free and bound variables in the formula for R so that a subset of the free variables in R are free in $\pi(R)$ and all others are bound. Certainly the hypothesis holds.

Case ($\sigma(R)$ where R is defined on schema (J,D)): Without loss of generality, we assume that the selection condition is a single theta comparison on a domain in the schema of R. The formula in the DRC expression for R is f which, by the induction hypothesis, has no nested negations, and the formula for the selection condition is a theta comparison $(X\ \theta\ Y)$, $(X\ \theta\ c)$ or $(c\ \theta\ X)$ where X and Y are variables for domains $D(j_1)$ and $D(j_2)$ and c is a constant drawn from $D(j_1)$. Then, the formula in the DRC for $\sigma(R)$ is $f \wedge (X\ \theta\ Y)$ or $f \wedge (X\ \theta\ c)$ or $f \wedge (c\ \theta\ X)$. If f is in DNF with no nested negations, then we know that we can distribute the conjunction across every disjunct in f without introducing any nested negations.

Case (($R \cup S$) where R and S are union compatible in the standard sense): If the formula for the DRC expression of R is f and the formula for the DRC expression of S is g, and by the induction hypothesis, f and g are in DNF with no nested negations when negations are pushed down, then with appropriate renaming and reordering, the formula for the DRC expression corresponding to $R \cup S$ is $f \vee g$. Because f and g are already in DNF, no further distribution is required. Certainly the disjunction of two formulas that satisfy the hypothesis will itself satisfy the hypothesis.

Case ($R \bowtie S$): The formulas for the DRC of R and S are the disjunctions $f_1 \vee \ldots \vee f_n$ and $g_1 \vee \ldots \vee g_m$ respectively, where any negated literals among the f_i's and g_j's are safe (i.e. bound) within each disjunct. Then with appropriate variable renaming and reordering, the formula for the DRC of $R \bowtie S$ is $f_1 \vee \ldots \vee f_n \wedge g_1 \vee \ldots \vee g_m$. After distribution, we have $f_1 \wedge g_1 \vee f_1 \wedge g_2 \vee \ldots \vee f_2 \wedge g_1 \vee \ldots \vee f_n \wedge g_m$ where each f_i and g_j is a conjunction of positive and negative literals so certainly the formula for the DRC of $R \bowtie S$ is also in DNF where the natural join does not introduce nested negations.

Case (($R - S$) where R and S are union compatible in the standard sense where the subtree for S has no difference operators): The formulas for the DRC of R and S are the disjunctions $f_1 \vee \dots \vee f_n$ and $g_1 \vee \dots \vee g_m$ respectively, where any negated literals among the f_i's are safe (i.e. bound) within each disjunct and there are no negated literals among the g_j's. The formula for the DRC of $R - S$ is then $f_1 \vee \dots \vee f_n \wedge \neg (g_1 \vee \dots \vee g_m)$. Distributing the negation across the disjuncts gives $f_1 \vee \dots \vee f_n \wedge \neg g_1 \wedge \dots \wedge \neg g_m)$ where each g_j is a conjunction of literals. Distributing the negated conjuncts across the f_i's gives $f_1 \wedge \neg g_1 \wedge \dots \wedge \neg g_m \vee f_2 \wedge \neg g_1 \dots . \vee f_n \wedge \neg g_1 \wedge \dots \wedge \neg g_m$. Some of the literals among the f_i's may be negated, but after pushing the negations into the g_j's and further distribution, into DNF, there is no introduction of nested negations.

Consequently, we conclude that for algebraic query expressions without a difference operator in the right-hand subtree of a difference operation, the formula in the corresponding DRC expression, when converted into DNF, will never encounter nested negations when pushing negations down to the literals. □

Knowing that such a relationship between algebraic expressions and DRC formulas holds, we can therefore establish that, for the subset of queries that limits the nesting of difference operators, the attribution constructed inductively in the algebra corresponds to the formal definition.

Theorem 5.4 The attribution algebra corresponds to the formal model where the nesting of difference operators is limited.
As with Lemma 5.1, we establish the theorem by induction on the number of operators in the algebraic expression, comparing the definitions constructed in the algebra to the formal definitions of the corresponding DRC equivalent. For notational convenience, all relations R, tuples t, and operators σ are implicitly extended.

In the base case of zero operators, the algebraic query expression is a single relation R on schema (J,D) or a constant relation. We know from Ullman (1988) that this is translated into an equivalent relational predicate $r(X_1,\dots,X_{max(J)})$ or a corresponding expression for the constant relation $\{t_1,\dots,t_n\}$ on (J,D) with formula $(X_1 = t_1D(1) \wedge \dots \wedge X_{max(J)} = t_1D(max(J))) \vee \dots \vee (t_nD(1) \wedge \dots \wedge t_nD(max(J)))$ where there is a disjunct for each tuple t_i.

For a base relation R on (J,D), we initialize the corresponding sets such that, for tuple $t \in R$, $t_C = R$ and for every j, $t_S[j] = R$, $t_I[j] = \varnothing$. Algebraically, then, for $t \in R$:
Comprehensive Attribution for a value $t_V[j]$ is $t_C = R$ for the expression $\pi_{D(j)}(\sigma_t(R))$;
Source attribution for a value $t_V[j]$ is $t_S[j] = R$ for the algebraic expression $\pi_{D(j)}(\sigma_t(R))$;
Relevant attribution for a value $t_V[j]$ is $<t_S[j] \cup t_I[j] = R$ for $\pi_{D(j)}(\sigma_t(R))$.

The corresponding formula for the equivalent DRC is just $r(X_1,\ldots,X_{max(J)})$ so for tuple $t \in R$, the comprehensive attribution for a value $X_i = c_i$ in t is the set of substitution lists $\{<c_1/X_1,\ldots,c_{max(J)}/X_{max(J)}>\}$ with no negated substitutions on the expression $\{X_i \mid \exists X_1,\ldots,X_{i-1},X_{i+1},\ldots X_{max(J)}\ r(X_1,\ldots,X_{max(J)}) \wedge X_1 = t_1 \wedge \ldots \wedge X_{max(J)} = t_{max(J)}\}$. Every substitution corresponds to the relation R, which is the attribution t_C in the algebra.

Likewise, the source substitution is just the substitutions in r corresponding to c_i/X_i with no negated substitutions on the same expression as for comprehensive substitution. But the source substitutions for c_i/X_i correspond only to the relation R, which is the attribution $t_S = R$ in the attribution algebra.

Finally, the relevant attribution in the base case is just the source substitution which corresponds to the algebraic definition $t_S[j] \cup t_I[j] = R$, and there are no negated predicates. Thus in the base case we confirm that the attribution algebra corresponds to the formal definitions of attribution.

In the inductive case, as with the relationship between the algebra and the DRC, we consider algebraic expressions with $n+1$ operators.

Lemma 5.5 Inductive case for comprehensive attribution

Case ($\pi(R)$): The DRC expression for the projection merely reassigns the set of free and bound variables in the formula for R so that a subset of the free variables in R are free in $\pi(R)$ and all others are bound. The projection of domains $K \subseteq J$ from scheme (J,D) so that the Comprehensive attribution for any tuple $t' \in \pi(R)$ is $\cup\, t_C\ \ \forall t \in R$ where $t[k] = t'[k]$ for all k (e.g. the weak duplicates t'). From the induction hypothesis we know that t_C corresponds to the substitutions in the equivalent DRC expression. The tuples t corresponding to a weak duplicate of t' are exactly those substitutions that agree in $t[k] = t'[k]$ and make the expression for R true. Therefore, any relation U in t_C corresponds to some substitution for a weak duplicate in the DRC expression for R. Thus we conclude, by the induction hypothesis, that the comprehensive attribution for a value in $\pi(R)$ corresponds to the formal definition.

Case ($\sigma(R)$ where R is defined on schema (J,D)): Without loss of generality, we assume that the selection condition is a single theta comparison on a domain in the schema of R. The algebraic comprehensive attribution for a value of $t' \in \sigma(R)$ is simply $t_C' = t_C$ for $t \in R$ and $\forall j$, $t'_V[j] = t_V[j]$. Likewise, because t' simply denotes the substitutions that make the formula in the expression for R true in addition to making the θ condition true, we know that the substitutions for $t \in R$ are the same substitutions for $t' \in R'$ so the algebraic definition corresponds to the formal model. Moreover, if there were any other substitutions $u \in R$ such that u satisfies θ and $u_V = t'_V$ then $t_V = u_V$ (or else R is not a relation). Thus, we conclude that the comprehensive attribution for a value in $\sigma(R)$ as computed by the attribution algebra corresponds to the formal definition.

Case (($R \cup S$) where R and S are union compatible in the standard sense): For a value in a tuple t that appears only in R or only in S then certainly the algebra and the formal definitions agree given the induction hypothesis that they agreed in R and in S. For a value in a tuple $t' \in R$ and $t' \in S$, the algebra will include t'_C from $R \cup t'_C$ from S. Likewise, the formula in the DRC is a disjunction $R \vee S$ and will include the substitutions from R and S corresponding to t'. By the induction hypothesis, the substitutions in S correspond to t'_C in S and the substitutions in R correspond to t'_C in R, therefore we conclude that the comprehensive attribution for a value in $R \cup S$ as computed by the attribution algebra corresponds to the formal definition.

Case ($R \bowtie S$): Where K from the select and then coalesce of R and S is empty, a value in a tuple t of $R \bowtie S$ comes either from R or from S but not both. If K is non-empty, then a value in a tuple t of $R \bowtie S$ could come from just R, just S, or both. However, regardless, the comprehensive attribution includes the relations in the comprehensive attribution of R and in the comprehensive attribution of S from the constituents for tuple t, r and s. Moreover, we know that there can only be one such $r \in R$ and $s \in S$ or R and S would not be relations. From the induction hypothesis, r_C and s_C correspond to the formal definition of the comprehensive attribution in R and S respectively. Therefore, every possible substitution that could produce r is reflected in r_C and likewise for s_C. Thus, though t may correspond to multiple permutations of disjunctions from the DRC for R and S, there are no permutations that are not captured in $r_C \cup s_C$, but this is the algebraic construction of the comprehensive attribution for a value in $t \in R \bowtie S$. Therefore, we conclude that the comprehensive attribution for a value in $R \bowtie S$ as computed by the attribution algebra corresponds to the formal definition.

Case (($R - S$) where R and S are union compatible in the standard sense and where the subtree for S has no difference operators): For a value in a tuple t of the difference where $t = r \in R$ and for which there is no s s.t. $r = s \in S$, the attribution algebra will return $r_C \cup_{\forall s \in S} s_C$. Note than any nested difference operators in R are captured in r_C while $\cup_{\forall s \in S} s_C$ captures the intuition of comparing every tuple of S to verify $r \notin S$. The corresponding DRC for R and S are formulas f and g in DNF so that $R - S$ is $f \wedge \neg g$. Distributing $\wedge \neg g$ over the disjuncts of f gives $f_1 \wedge \neg g \vee f_2 \wedge \neg g \vee \ldots \vee f_n \wedge \neg g$. For tuple $t = r \in R$, r_C corresponds to the substitutions in $f_1 \ldots f_n$ such that $t = r$ makes f_i true by the induction hypothesis. Likewise, $\cup_{\forall s \in S} s_C$ corresponds to the set of all substitutions that makes g true. Thus we conclude that the comprehensive attribution for a value in $R - S$ as computed by the attribution algebra corresponds to the formal definition. □

Lemma 5.6 Inductive case for source attribution

Case ($\pi(R)$): Assume $\pi(R)$ is on scheme (J,D) for function p. The DRC expression for the projection merely reassigns the set of free and bound variables in the formula for R so that a subset of the free variables in R are free in $\pi(R)$ and all others are bound. From the induction hypothesis we know that $t_S \in R$ corresponds to the source substitutions in the equivalent DRC expression. The tuples $t \in R$ that produce the weak duplicate $t' \in \pi(R)$ are exactly those substitutions that agree in $t[p(j)] = t'[j]$ and make the DRC expression for R true. Therefore, $\forall j$, any relation U in the set $t_S[j]$ corresponds to some substitution for a weak duplicate in the DRC expression for R. Thus we conclude, by the induction hypothesis, that the source attribution for a value in $\pi(R)$ corresponds to the formal definition.

Case ($\sigma(R)$ where R is defined on schema (J,D)): Without loss of generality, we assume that the selection condition is a single theta comparison on a domain in the schema of R. The algebraic source attribution for a value of $t' \in \sigma(R)$ is simply $t'_S = t_S$ for $t \in R$ and $\forall j$, $t'_V[j] = t_V[j]$. Likewise, because t' simply denotes the substitutions for the free variables in the expression for R such that both R and and the θ condition are true, we know that the substitutions for $t \in R$ are the same substitutions for $t' \in R'$ so the algebraic definition corresponds to the formal model. Moreover, if there were any other substitutions $u \in R$ such that u satisfies θ and $u_V = t'_V$ then $t_V = u_V$ (or else R is not a relation). Thus, we conclude that the source attribution for a value in $\sigma(R)$ as computed by the attribution algebra corresponds to the formal definition.

Case ($(R \cup S)$ where R and S are union compatible in the standard sense): For a value in a tuple t that appears only in R or only in S then certainly the algebra and the formal definitions agree given the induction hypothesis that they agreed in R and in S. For a value in a tuple $t' \in R$ and $t' \in S$, the algebra will include $t'_S \in R \cup t'_S \in S$. Likewise, the formula in the DRC is a disjunction $R \vee S$ and will include the substitutions from R and S corresponding to t'. By the induction hypothesis, the substitutions in S correspond to t'_S in S and the substitutions in R correspond to t'_S in R, therefore we conclude that the source attribution for a value in $R \cup S$ as computed by the attribution algebra corresponds to the formal definition.

Case ($R \bowtie S$): Where K from the select and then coalesce of R and S is empty, a value in a tuple t of $R \bowtie S$ comes either from R or from S but not both. If K is non-empty, then a value in a tuple t of $R \bowtie S$ could come from just R, just S, or both. Consider the case where the value in t, $t_V[j]$ comes from $r \in R$ or $s \in S$ but not both. First, for any tuple t, we know that there can only be one such r and one such s. From the induction hypothesis, if K is empty or the value does not come from $D_1(k) = D_2(k)$, then it is easy to see that $t_S[j]$ must either be equal to some $r_S[j_1]$ or some $s_S[j_2]$ where R and S are defined on (J_1,D_1) and (J_2,D_2) respectively. If the value does come from some $D_1(k) = D_2(k)$, then algebraically, we know that $t_S[j] = r_S[k] \cup s_S[k]$. In the equivalent formula of the DRC where K is non empty, we know that variable renaming and reordering results in multiple occurences of the same variable name in predicates of R and predicates of S. But every substitution must correspond to predicates of R in $r_S[k]$ and a predicates of S in $s_S[k]$ and none others by the induction hypothesis. Then the source substitutions in the formal model correspond to the algebraic source substitution and we conclude that the source attribution for a value in $R \bowtie S$ as computed by the attribution algebra corresponds to the formal definition.

Extended algebra

Case (($R - S$) where R and S are union compatible in the standard sense where the subtree for S has no difference operators): For a value in a tuple t of the difference where $t = r \in R$ and for which there is no s s.t. $r = s \in S$, the attribution algebra will return r_S. The corresponding DRC for R and S are formulas f and g in DNF so that $R - S$ is $f \wedge \neg g$. For tuple $t = r \in R$, r_S corresponds to the substitutions in $f_1 \ldots f_n$ such that $t = r$ makes f_i true by the induction hypothesis. Likewise, the tuple t should not appear in any disjunct of g therefore no substitutions of g should appear as a source for values of t. Thus we conclude that the source attribution for a value in $R - S$ as computed by the attribution algebra corresponds to the formal definition. □

Lemma 5.7 Inductive case for relevant attribution

Case ($(\pi(R))$): In the algebra, we project the domains $D_2 \subseteq D_1$ from schema (J_1, D_1). From the induction hypothesis, we know that for any tuple $t \in R$, $\forall j$, $t_S[j] \cup t_I[j]$ returns the set of relation names that contain the substitutions returned by $Relevant(D_1(j))$ in the DRC. Likewise, we know that the DRC for $(\pi(R))$ simply reassigns the free and bound variables in the formula for the expression, which means that in the formal model, the expression is the same so $Relevant(D_2(j_2)) = Relevant(D_1(p(j_2)))$. Thus $\forall t' \in \pi(R)$, the relevant substitutions in the DRC are the same as that for R corresponding to the algebraic definition where $t'_S[j_2] = t_S[p(j_2)]$ and $t'_I[j_2] = t_I[p(j_2)]$. Weak duplicates are simply those substitutions that agree in all of the values of j_2 but not all the values of j_1. But the formal model is a set of substitutions, so for any instance corresponding to the free variables, the substitution is the set of all substitutions that make one instance true and is just the set of all weak duplicates. In the algebra, this is the union of $t'_S[j_2]$ and $t'_I[j_2]$ over all t' that agree in the values $t'_V[j]$.

Case ($\sigma(R)$ where R is defined on schema (J,D)): Without loss of generality, we assume that the selection condition is a binary theta comparison θ, K on a domain in the schema of R. As noted in the definitions earlier, for simplicity, we invoke a function $Relevant(K)$ to return the same domain variables in the algebra as in the DRC expression. Therefore, by the equivalence of $Relevant(X)$ where X ranges over the domain variables in the DRC expression and $Relevant(D(j))$, we see that the algebra begins with the initial relevant relations (induction hypothesis) and incorporates only those relations containing any domain variable X. Hence, we conclude that for $t' \in \sigma(R)$, $\forall j \in J$, the relevant attribution for $t'_V[j]$ corresponds to the relevant substitutions for the set of free variables on the same domain $D(j)$ in the DRC.

Case (($R \cup S$) where R and S are union compatible in the standard sense): For a value in a tuple t that appears only in R or only in S then certainly the algebra and the formal definitions agree given the induction hypothesis that they agreed in R and in S. For a value in a tuple $t \in R$ and $t \in S$, the algebra will combine t_S from $R \cup t_S$ from S and treat the t_I sets similarly (see: weak duplicate elimination). Likewise, the formula in the DRC is a disjunction $R \vee S$ and will include the relevant substitutions from R and S corresponding to the free variables as they appear in relational predicates R and S. By the induction hypothesis, the relevant substitutions in S correspond to $t_S \cup t_I$ in S and the substitutions in R correspond to $t_S \cup t_I$ in R, therefore we conclude that the relevant attribution for a value in $R \cup S$ as computed by the attribution algebra corresponds to the formal definition.

Case ($R \bowtie S$): As in other proofs for natural join, we rely here upon composition and the fact that natural join is defined as a Cartesian product followed by a selection, a coalesce, and a projection. We show that the property holds for natural join with no join variables (Cartesian product), and then rely upon the proofs for selection and projection shown earlier.[34] Every tuple of $R \bowtie S$ is comprised of a tuple $t_1 \in R$ and a $t_2 \in S$. From the induction hypothesis, we know that $\forall j$, *Relevant* in t_1 and *Relevant* in t_2 contains the relations for the substitutions in the corresponding relational predicates of the DRC expression. In concatenating a tuple of R and a tuple of S, certainly the property still holds. Thus, we may continue to apply the induction hypothesis to the subsequent selection on equality and finally project out redundant attributes.[35]

Case (($R - S$) where R and S are union compatible in the standard sense where the subtree for S has no difference operators): For a value $t_V[j]$ in a tuple t of the difference where $t = r \in R$ and for which there is no s s.t. $r = s \in S$, the attribution algebra will return $t_S[j] \cup t_I[j]$ where $t_I[j] = r_I[j] \cup \forall s \in S\ s_C$. In particular, every tuple $s \in S$ becomes relevant because it is used to verify that the instance $t_V[j]$ (defined as the tuple of R containing $t_V[j]$) does not appear in S. r_I is how the substitutions from nested difference operators are carried forward. In Chapter 4 we spoke of the additivity property in negation and we see the importance here. We account for nested difference operators in the left hand side (minuend) of a difference operator by continuing to add to t_I. Thus we conclude that the relevant attribution for a value in $R - S$ as computed by the attribution algebra corresponds to the formal definition. □

Hence from the base case and Lemmas 5.5 through 5.7, we conclude that when we do not allow nesting of difference operators in the left-hand side of a difference operator, the attribution algebra corresponds to the formal model. □

Particularly interesting about the limitations that we impose on the difference operator is that for such algebraic expressions, the corresponding DRC corresponds to the subset of DRC expressions for which composition holds. Therefore, while the algebra constructs attribution inductively from the leaves of the query tree up to the root, we are equally assured that we can compose attribution by beginning at the root and drilling down to the base relations at the leaves.

[34] Note that because we use the function *Relevant* defined to match the formal model in our definition of *extended select* and then explicitly select on equality, the selection variables are by definition relevant to one another and thereby implicitly coalesce the relevant (intermediate) sets.

[35] The reader may recall that in proving the closure of the extended relational algebra, we verified that the result of a Cartesian product on extended relations is an extended relation.

5.5 Summary

In this Chapter, we have presented an extension to the relational algebra that inductively constructs the attribution for value-level result granules in an eager manner, as a part of query processing. Mindful of the potential explosion in the amount of attribution metadata that such a process can create, the algebra manages source granules at the relation level.

We first formalize the relationship between the standard relational algebra and the extended algebra. Subject to some restrictions on the use of negation in query expressions, we then establish that the attribution generated by the extended algebra does correspond to the formal definition as established in Chapter 4. The relationship between the composition property of attribution and the inductive algebraic process suggests some interesting possibilities for deploying attribution as an accompaniment to a standard query processor or as an external, network service for lazy attribution processing. Moreover, the parameterization of attribution characteristics in the algebra hints at the potential for incorporating either other types of metadata or more complex functions (e.g. data quality) of existing attribution characteristics. We return to these issues in the Conclusion.

6 Attribution and the Web

We began this thesis by hypothesizing an imaginary on-line travel resource integrator that could answer queries not only based upon its own knowledge but also by possibly gathering and utilizing information from any number of unknown sources. Such systems, however, are no longer hypothetical. Integration, whether for travel, finance, healthcare, current events, etc. is now a trademark application of the World Wide Web.

We saw in Chapter 1 how attribution may serve many different roles in data integration. As a consequence, we identified several dimensions to describe the problem of attribution. Although our initial interest in this thesis stemmed from the Web, Web querying is an active research topic that has only recently begun to approach a uniform standard (Chamberlin et al. 2001a; Fernandez and Marsh 2001). Like the integration that it enables, the underlying theory of Web querying combines several intellectual disciplines including databases, information retrieval, and library science (deBakker and Widarto 2001; Katz 2001; Lenz 2001). As a consequence, we simplified our task by casting the problem of attribution in the context of the relational data model. We presented the formal model in Chapter 4.

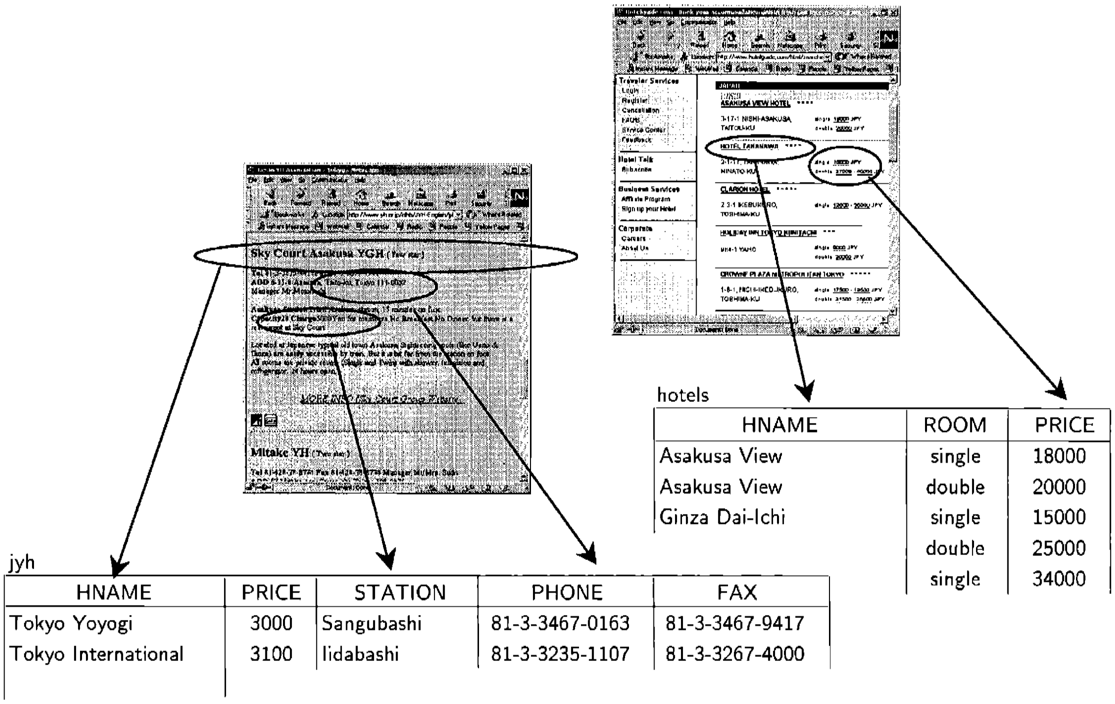
In this Chapter, we return to the Web. Specifically, we consider how our formal model, developed in the context of the relational data model, relates to the semistructured data model of the World Wide Web. We begin with a very brief overview of some general, semistructured data concepts. Next, we consider how our attribution intuitions from Chapter 3 relate to the semistructured space. Finally, we consider limitations of applying our formal model of attribution to the Web, referring the reader to work by Buneman et al (2001; 1998; 2000; 2001) on attribution (provenance) for semistructured data.

6.1 Semistructured data models

Research on semistructured data is often confused with evolution of the Web. However, the challenge of data integration existed long before the Web. Current work on semistructured data borrows from portions of the database literature that is often implicitly associated with Web querying: Tsimmis, LORE, Infomaster, Information Manifold (Abiteboul et al. 1997; Chawathe et al. 1994; Duschka and Genesereth 1997a; Duschka and Genesereth 1997b; Levy, Rajaraman, and Ordille 1996). Despite their clear applicability to data on the Web, however, these works were all pursued in the general context of data integration. Indeed, from a data integration perspective, the Web has represented a working infrastructure that simultaneously emphasized the need for and provided a testbed for research on integration and semistructured data (Buneman 1997) (Florescu, Levy, and Mendelzon 1998). In the past five to ten years, interest in and research on semistructured data has exploded. Our goal here is not to summarize the field. Others have covered the foundations (Abiteboul, Buneman, and Suciu 2000). Our goal, instead, is to touch on enough of the basic principles to inform a discussion of how attribution principles might apply in a semistructured environment.

6.1.1 Semistructured data representation

Research in semistructured data models is driven, in no small part, by the observation that data in the "real world" seldom conforms to the well-behaved assumptions that underlie the relational data model. In particular, while data may often be arranged to have the same appearance, the underlying structure or schemas can differ significantly. Consider, for example, the Travel Resource Integrator from Chapter 1. The travel examples used throughout Part 1 of this thesis draw data from a number of on-line, Web-accessible travel guides. As indicated in Figure 6.1, our initial intuition was to model the data from these Web travel guides as the relations of Chapter 3.



hotels

HNAME	ROOM	PRICE
Asakusa View	single	18000
Asakusa View	double	20000
Ginza Dai-Ichi	single	15000
	double	25000
	single	34000

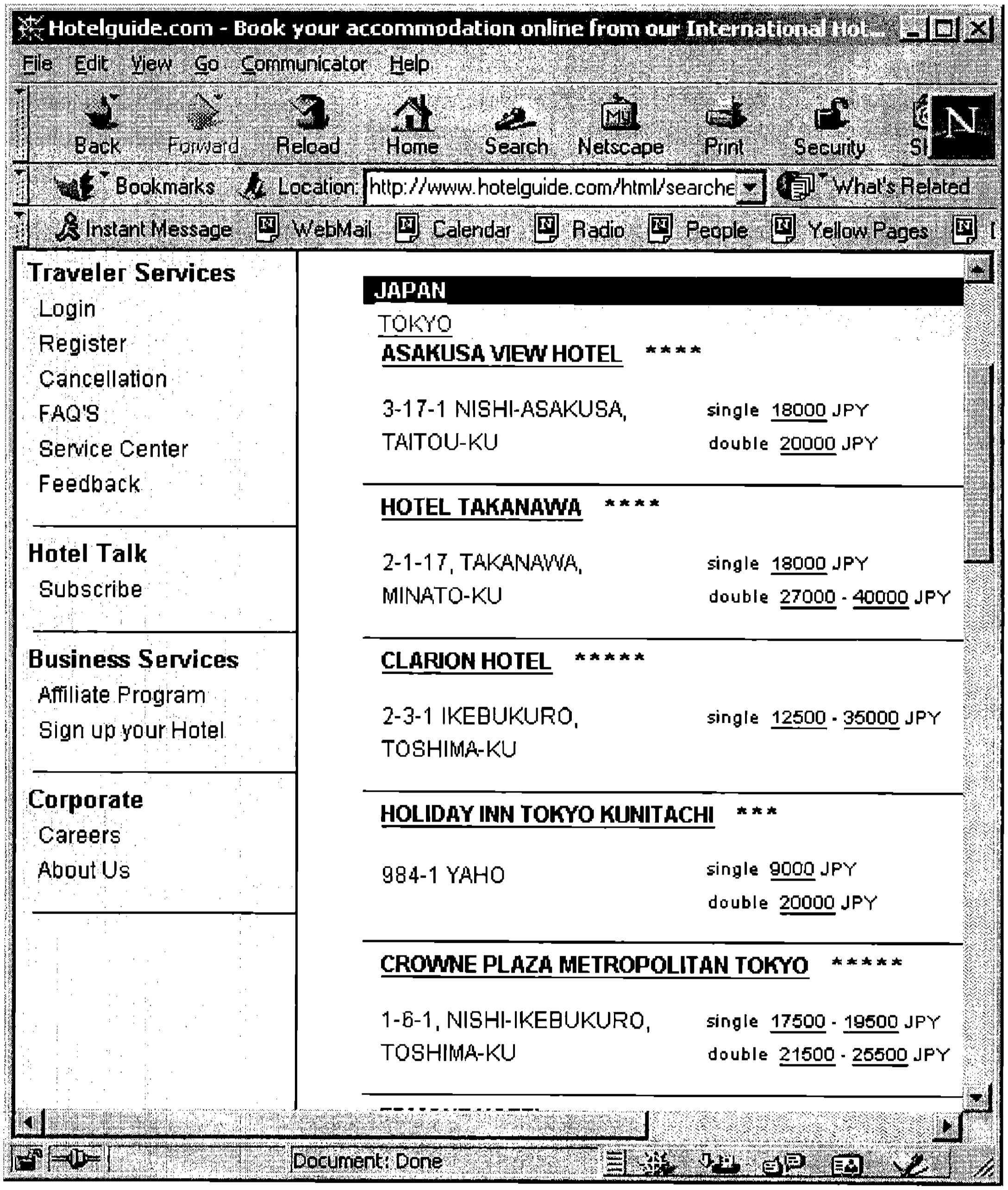
jyh

HNAME	PRICE	STATION	PHONE	FAX
Tokyo Yoyogi	3000	Sangubashi	81-3-3467-0163	81-3-3467-9417
Tokyo International	3100	Iidabashi	81-3-3235-1107	81-3-3267-4000

Figure 6.1 The Web as a relation

ATTRIBUTION AND THE WEB

Upon closer inspection, however, it quickly becomes apparent that the relational perenity which we assumed in Chapter 3 breaks down. Consider the Web guide "The Hotel Guide" from which we populated the relation table "hotels" (hotelguide.com 2001). We include one page of hotels in Tokyo, Japan from hotelguide.com in Figure 6.2. Aside from the fact that there are a number of hotels that we omitted simply for tractability reasons, we quickly notice that there are some inconsistencies. Not all hotel listings match the entry for the


Hotelguide.com - Book your accommodation online from our International Hot...
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Sh
Bookmarks Location: http://www.hotelguide.com/html/searche What's Related
Instant Message WebMail Calendar Radio People Yellow Pages
Traveler Services
Login
Register
Cancellation
FAQ'S
Service Center
Feedback
Hotel Talk
Subscribe
Business Services
Affiliate Program
Sign up your Hotel
Corporate
Careers
About Us
Document: Done

JAPAN

TOKYO

ASAKUSA VIEW HOTEL ****

3-17-1 NISHI-ASAKUSA, TAITOU-KU — single 18000 JPY, double 20000 JPY

HOTEL TAKANAWA ****

2-1-17, TAKANAWA, MINATO-KU — single 18000 JPY, double 27000 - 40000 JPY

CLARION HOTEL *****

2-3-1 IKEBUKURO, TOSHIMA-KU — single 12500 - 35000 JPY

HOLIDAY INN TOKYO KUNITACHI ***

984-1 YAHO — single 9000 JPY, double 20000 JPY

CROWNE PLAZA METROPOLITAN TOKYO *****

1-6-1, NISHI-IKEBUKURO, TOSHIMA-KU — single 17500 - 19500 JPY, double 21500 - 25500 JPY

Figure 6.2 Hotels in Tokyo, Japan found in www.hotelguide.com

"Asakusa View Hotel". Some entries, like the "Clarion Hotel" may not quote a price for doubles. Others, like the "Hotel Takanawa" may actually indicate a range of prices by listing two values for a "single". Were the hotelguide.com to treat hotel entries as tuples in a relation, the schema might include the union of all schema elements and set missing values to NULL. Rather than treat these values as explicitly NULL, they are instead simply non-existent. There are certainly other ways in which data on the Web does not conform to the relational model (Florescu, Levy, and Mendelzon 1998). However, our goal here is to motivate the "schema-less" or "self-describing" property that is characteristic of all semistructured data models.

Though there are multiple approaches to semistructured data representations, a common theme in the different representations is an explicit rendering of label-value pairs as a generalization of the "attribute-value" pairs in relations. By explicitly encoding every value with a label, semistructured data models carry structure as a part of the data rather than associating tuples (lists of values) with some external schema that conveys structure and meaning.

The concept of self-describing data is perhaps most easily conveyed in a tree or graph. In this overview, we follow the literature by describing the basic model as an edge-labeled graph where edges one of two categories of information. First, edges may contain typed-data commonly associated with the values in the attribute-value parlance of the relational data model. Second, edges may contain names or scalars that are colloquially associated 7with the "attribute" of an attribute-value pair.

In Figure 6.3, we suggest a semistructured model for two hotel entries from hotelguide.com. The reader should notice how every label or edge is associated with a value where the value may be a data value or a node denoting a set of label value pairs.

Although not depicted here, the basic model for semistructured data allows for the explicit association of a unique identifier with a node in the graph. Object identity provides a convenient mechanism for extending tree-structures, such as those depicted in Figure 6.3, into a graph.[36] The reader may also notice that in our example, there are no values on internal edges. Though not necessary, the basic model does not allow values on internal edges. Whether values are assigned to nodes versus edges and whether values are allowed on internal nodes or edges are all variations on the basic model.[37]

Following (Abiteboul, Buneman, and Suciu 2000), we can serialize our graph using the following grammar. If *s* is a semistructured data expression and *oid* is an object identifier that names a node from which edge(s) depart:

[36] Our existing hotel data might not provide the best opportunity for demonstrating graph extensions. The reader is encouraged to refer to (Abiteboul, Buneman, and Suciu 2000) for examples. The reader may also be familiar with the use of IDREF in XML to serve a similar function (Bray, Paoli, and Sperberg-McQueen 1997).

[37] The reader is encouraged to see (Abiteboul, Buneman, and Suciu 2000) for a discussion of these variations.

<s> := *<value>* | oid *<value>* | oid
<value> := atomic value | *<complex value>*
<complex value> := {label: *<s>*,..., label: *<s>*}

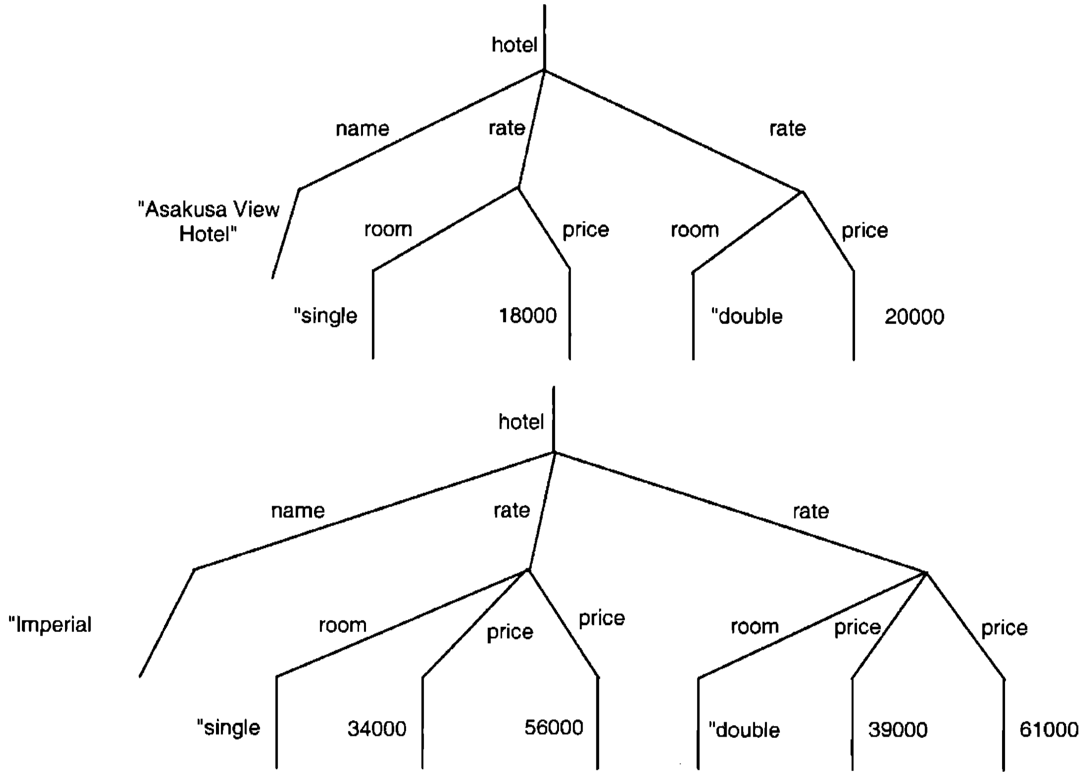


Figure 6.3 Semistructured data from www.hotelguide.com

Example 6.1 Serializing a graph
If we follow convention and name *oid*s using ampersands (e.g. &o1), we can serialize Figure 6.3 as follows:

```
{hotel:  &o1{name:  &o2"Asakusa View",
             rate:  &o3{room:  &o4 "single",
                        price: &o5 18000}
             rate:  &o6{room:  &o7 "double",
                        price: &o8 20000}
            },
...
 hotel:  {name:  "Imperial Hotel",
             rate:  {room:  "single",
                     price:  34000,
                     price:  56000}
             rate:  {room:  "double",
                     price:  39000,
                     price:  61000}
```

```
                },
   ...
} □
```

In our serialization of Figure 6.3, we deliberately omitted object identifiers from the second hotel listing. We did so to emphasize the characteristic that, like object-oriented models in general, the basic semistructured data model supports node identity. The model allows for the explicit assignment of a unique identifier to a node. In the absence of assignment, every node has an implicit identifier to establish the uniqueness used in data processing.

6.1.2 Semistructured data manipulation

Query languages serve two fundamental objectives: selection (to avoid confusion with the relational select (σ) operator, we may also use the term "extraction") and presentation. A relational query operator takes one or more relations, each of which is defined on a schema, and extracts some subset of tuples. A new relation is constructed from the extracts. Similarly, operators to manipulate semistructured data take, as arguments, the nodes and edges that constitute one or more graphs. After extracting some subset of nodes (and edges), a semistructured operator constructs a new graph. Just as there are different relational query languages, there are different semistructured query languages. In this subsection, we focus on a few shared concepts for selecting and presenting semistructured data.

6.1.2.1 Data extraction

All semistructured query languages support an elementary form of extraction based upon path expressions. Path expressions are the basic construct with which semistructured query languages specify nodes in a graph. A path is a well-understood concept from graph theory, but we can define a path on semistructured data informally as a sequence of edges between two nodes. The path expression $/l_1/l_2/\ldots/l_n/l_b$ denotes a path from node a to node b if the graph contains nodes $x_1 \ldots x_n$ and edges such that $(a\ l_1\ x_1)$, $(x_1\ l_2\ x_2), \ldots,$ $(x_n\ l_b\ b)$ (Abiteboul, Buneman, and Suciu 2000). We may then think of a path expression as a query constructor. The result of a path expression applied to a graph is the set of all nodes b for which there are edges l_1, $l_{2, \ldots}$ l_n, l_b from a to b.

Example 6.2 Path expressions

For example, the path expression `/hotel/name` applied to the graph of Example 6.1 returns the set of nodes for the edges `"Imperial Hotel", "Asakusa View", etc.`

The path expression /hotel/rate/price returns the set of nodes for the edges `18000, 20000, 34000, 56000, 39000, 61000, etc.` □

Path expressions are richer than a sequence of labels, however. By applying regular expressions on the alphabet of edge labels, we expand the paths denoted by (and hence the set of nodes returned by) a single path expression.

Example 6.3 Regular expressions in path expressions
Following the regular expression syntax in Perl, we may write the following path expression: `/(hotel|hostel)/*/price`. Certainly the path: `/hotel/rate/price` matches the pattern of the path expression; among others, the path expression returns the set of nodes for all hotel prices from Figure 6.3. We could also imagine integrating data from the jyh relation with data from www.hotelguide.com by expanding the graph of Figure 6.3 with hostel edges of the form seen in Figure 6.4. Now our path expression also matches the path `/hostel/charge/member/price`. The set of nodes returned by the original path expression now also includes nodes associated with hostel prices. □

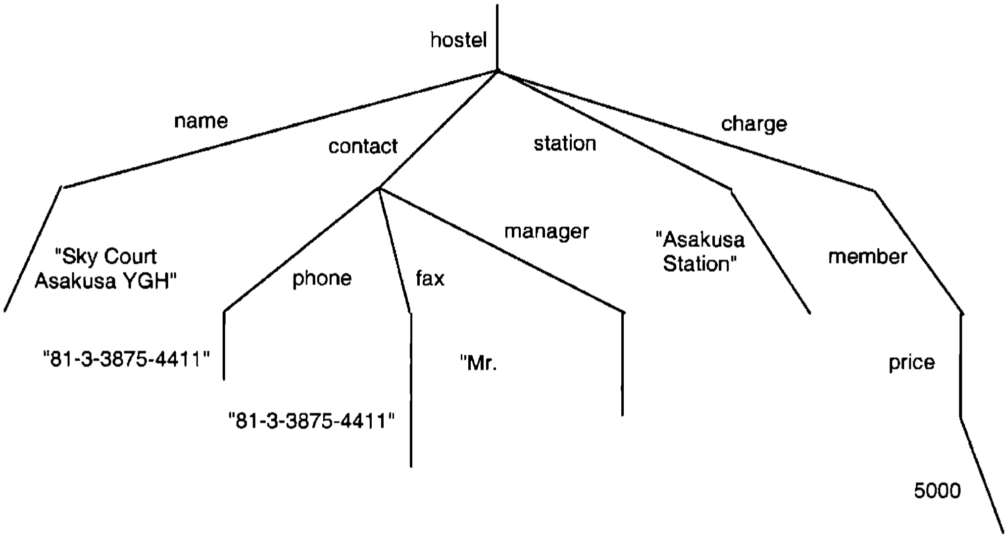


Figure 6.4 Representing hostel Web data in a graph

6.1.2.2 Data presentation

While path expressions return a set of nodes, as a query language, path expressions are incomplete. Path expressions can extract, but a set of nodes does not by itself constitute a graph.[38] We need tools to control presentation (i.e. construct a graph from the nodes in the result of a path expression). The use of variables, in conjunction with path expressions, supports presentation. The result of a path expression is assigned to a variable. These variables are used in the specification of an output path. The output path is a template for the graph of the result of a path expression. In the same way, the "select" clause of an SQL

[38] The closure property suggests that, given graphs on inputs, the query language returns a graph.

statement defines the schema of the result. Variables and path expressions together complete the basic elements of a semistructured query language. Details of explicit query syntax may vary among specific semistructured query languages, but the roles served by variables and path expressions are roughly the same.

Example 6.4 Constructing the result graph of a semistructured query
We use the same path expression as before to extract possible prices for lodging in and around Tokyo, Japan except now we assign node instances to the variable X:
`/(hotel|hostel)/*/price X`. Now we build a path as a template for the output of the path expression: `/lodging/price/X`. This path corresponds to the graph of Figure 6.5. □

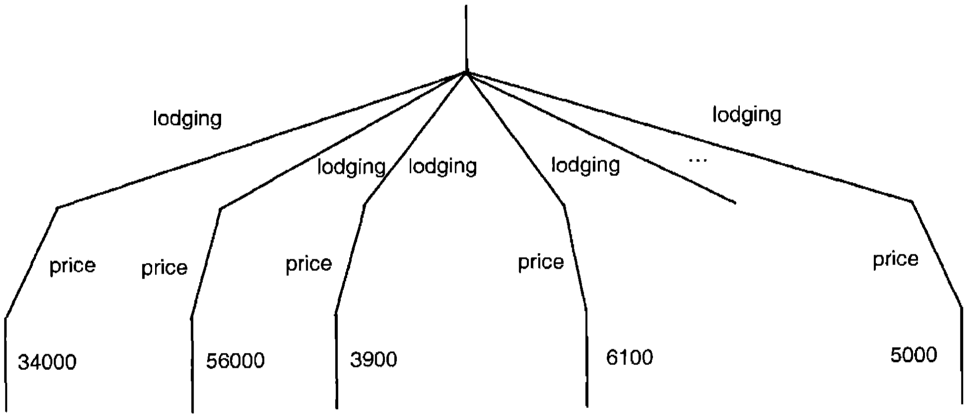


Figure 6.5 Semistructured query result

6.1.2.3 Extending data manipulation capabilities

While path expressions and variables provide the basic infrastructure for a rudimentary, semistructured query language, these tools also support much richer classes of queries. With variable assignment, semistructured languages can support θ-comparisons to further restrict the subset of nodes extracted. Through variable assignments and nested queries, we can support complex graph restructuring.

Example 6.5 θ-comparisons and graph restructuring in semistructured queries
Our query might first consider each hotel or hostel separately.

```
/(hotel|hostel)/ X
```

A "for-each" conjunction of conditions on every `X` nests one query within another. For each hotel or hostel node, we assign the name to `Y` and the price to `Z`.

```
/X/name Y
/X/*/price Z
```

We can apply a boolean test on prices to further restrict nodes in the result graph.

```
Z < 35,000
```

We then use our variables to define a path as a template for the result graph.

```
/Y/price/Z
```

The final result graph is depicted in Figure 6.6. □

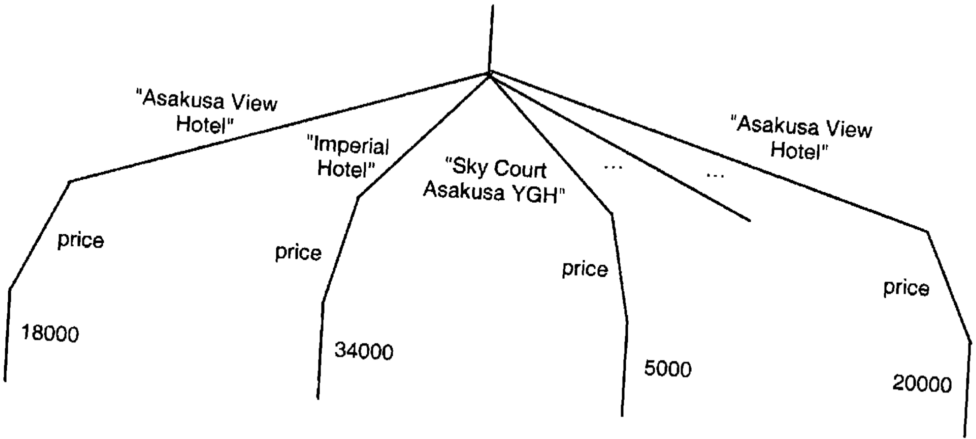


Figure 6.6 Nested queries and complex restructuring

In introducing semistructured data manipulation, we have deliberately left out many details that we feel are less germane to attribution. Most semi-structured query languages use SQL-like `select-from-where` syntax and some familiar notation for expressing regular expressions on paths.[39] In addition, we can apply regular expressions on labels themselves. For example the conjunction of two path expressions `/(hotel|hostel) X` and `/X/name/"A*"` represent a pattern to get all lodging nodes with names beginning with the letter "A." Other issues involve duplicate management in the face of object identity and the type coercion required to perform θ-comparisons on edge labels or restructure graphs using internal edge labels as values and vice versa. The reader is encouraged to consult other sources on the subject (Abiteboul 1997; Abiteboul, Buneman, and Suciu 2000; Abiteboul et al. 1997; Abiteboul and Vianu 1997; Buneman 1997; Buneman et al. 1997; Buneman, Deutsch, and Tan 1998; Chawathe, Abiteboul, and Widom 1999; Fernandez et al. 1997a; Fernandez et al. 1997b; Lenz 2001).[40]

6.2 Attribution intuitions and semistructured data

Having introduced some of the basic principles of semistructured data representation and manipulation, we next consider how some of our attribution intuitions apply to the semistructured context. While the relationship between attribution in the different models is imperfect, in this first section, we consider only how the intuitions do match. In the following section we raise some of the complications.

[39] Familiar notations for paths include "." or "/" separators. Regular expression symbols include "*" for zero or more, "?" for zero or one, "+" for one or more, etc.

[40] We intentionally steered away from explicit reference to XQuery, XPath, and other rapidly evolving World Wide Web Consortium (W3C) standards for querying XML. We did so first to avoid the popular misconception that XML queries are synonymous with rather than simply one (albeit prominent) example of semistructured querying. Second the W3C standards were evolving too rapidly for us to consistently track in this document. We do include references to the W3C work both in the in-text citations above and in the References.

Our general intuition in the formal model of attribution was of substitutions that make an expression true. If we think of a semistructured query as a conjunction of path expressions, the analogy seems simple enough. The attribution for a semistructured query constitutes the subgraphs that match a particular pattern corresponding to the nodes in a result graph.

Example 6.6 Subgraphs that match a particular pattern
In Example 6.3, we gave the following path expression: `/(hotel|hostel)/*/price`. Based upon our sample data from Figures 6.3 and 6.4, we know that the following paths all match the pattern:

```
/hotel/rate/price for the nodes with:
/hotel/name/"Asakusa View" and /hotel/rate/price/18000;
/hotel/name/"Asakusa View" and /hotel/rate/price/20000;
/hotel/name/"Imperial Hotel" and /hotel/rate/price/34000;
/hotel/name/"Imperial Hotel" and /hotel/rate/price/56000;
/hotel/name/"Imperial Hotel" and /hotel/rate/price/39000;
/hotel/name/"Imperial Hotel" and /hotel/rate/price/61000;
```

and

```
/hostel/charge/member/price for the node with
/hostel/name/"Sky Court Asakusa YGH" and /hostel/charge/member/price/5000
```

□

In the formal model, we explored different categories of equivalences. For the concept of strict equivalence, the difference between object-identity and value-equivalence introduces a slight inconsistency, but even with object-identity, we can imagine multiple paths in a graph to the same node.

Example 6.7 Strict equivalence: multiple paths to the same node in a graph
For example, suppose two different youth hostels shared the same manager. We illustrate such a possibility in Figure 6.7. □

The potential for cycles in a graph, of course, will also result in multiple paths to the same node. In the formal model we encountered a related problem posed by the potential introduction of redundant conjuncts. The relational calculus has the concept of a minimal query and the question of a minimal path is an open question that we raise as a challenge below and direct the reader to external references (Abiteboul, Hull, and Vianu 1995).

Equivalence through composition is a second category of equivalence. In the formal model, attribution composition stems from query composition (i.e. using the result of one query as the input to another as in IDB) The principle behind attribution composition is to recursively construct attribution in a step-wise fashion rather than to unfold the entire query a'priori or to carry metadata attribution forward with each value, updating with every additional operator.

Query and attribution composition has particular relevance for semistructured data and the Web in particular. Querying against one or more graphs returns a new graph that itself can serve as a source for a new path expression. Web portals and other aggregation engines serve in this very manner. In Chapter 1, we recounted the lawsuit between Priceman and MySimon. We may characterize a page in MySimon as the result of query that itself became a source for Priceman. Analogously, we may compose attribution in a stepwise fashion.

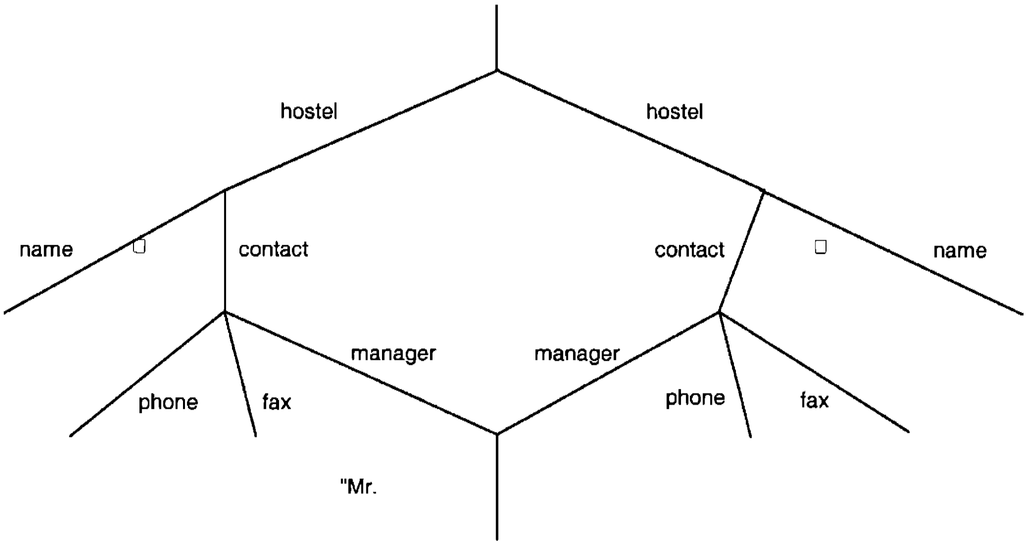


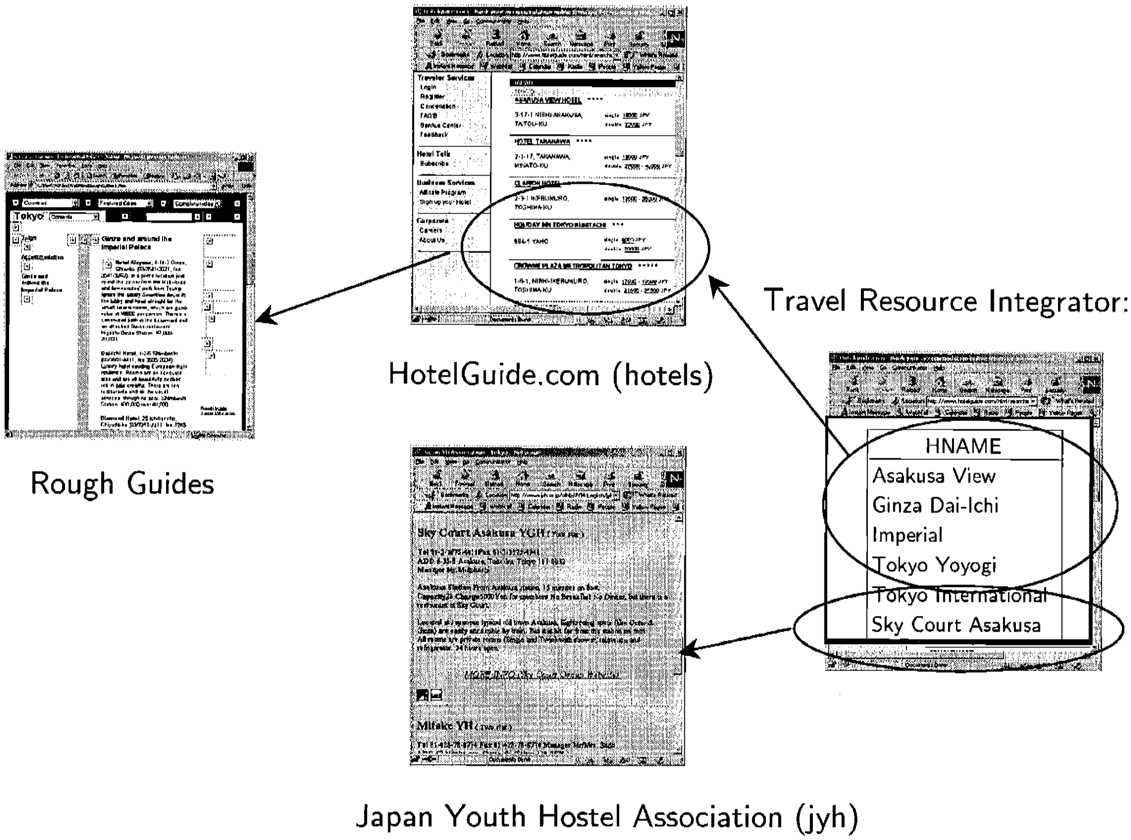
Figure 6.7 Strict equivalence in semistructured data

Example 6.8 Attribution composition for semistructured data
From our Travel Resource Integrator of Chapter 1, we could imagine attributing the result of a query on Tokyo sites to sources including www.hotelguide.com and www.jyh.com. We could equally imagine that these sites might in turn aggregate information from additional sources. Perhaps we might attribute the "Asakusa View Hotel" in www.hotelguide.com and discover that the listing was itself extracted from a RoughGuides travel guide as in Figure 6.8.[41] □

Finally, we consider our observations from Chapter 4 on coarse- and fine-grained source and result granularity. Our intuition for result granularity was the thought of rolling-up attribution from a value to its identifying tuple or to its domain. Likewise, a domain or a tuple may share attribution characteristics with the containing relation. Attribution at a higher level of result granularity aggregates the attribution for each constituent. Source granularity combines our ideas about result granularity and composition. Recognizing that a result granule associates attribution with some subset of values, and that the result granule can itself constitute a source for a composed query, we arrive at the concept of a source granule. Rather than attributing from substitutions in a source relation, we might attribute to source tuples or source relations.

[41] hotelguide.com does not indicate that it uses Baedekker's as an external reference and we use the example here merely to illustrate the concept of query and then attribution composition.

In Example 6.5, we saw how we could use a path expression to reference an internal node. As with our example, at least some semistructured query languages use references to internal nodes as a form of syntactic sugar for nesting queries on the independently named sub-elements (Abiteboul, Buneman, and Suciu 2000). Accordingly, we might envision using this notation to associate attribution with some internal node, implicitly referencing the subgraph rooted at the internal node. Attribution to an internal node would correspond to the idea that coarse granularity captures the attribution for each constituent. Moreover, because path expressions constitute query selection constructors, we can think of specifying arbitrary granules with query expressions. Colloquially, we can talk about attributing parts of a Web page rather than the page en masse as in a bibliography or individual items as in a footnote. Indeed it was because of observations about granules in semistructured data that we sought an analogy in the relational context.


Travel Resource Integrator:
HotelGuide.com (hotels)
Rough Guides
Japan Youth Hostel Association (jyh)

HNAME
Asakusa View
Ginza Dai-Ichi
Imperial
Tokyo Yoyogi
~~Tokyo International~~
Sky Court Asakusa

Figure 6.8 Attribution composition in semistructured queries

Example 6.9 Granularity for semistructured data attribution
Referring again to Figure 6.8, we indicate how a result may be separated into different granules. Hotel information comes only from HotelGuide.com and likewise for hostel information. Similarly, we may not have used all of the information from HotelGuide.com,

so we can separate their data into source granules. If information about a hostel's address information comes from a different place than pricing and management information, then we may think of each source as the result of a query on some other source and attribute accordingly. □

6.3 Challenges for attributing the Web

While many of our intuitions appear to map in a straightforward manner to the Web, there are a number of confounding factors that make attribution on the Web a challenge in its own right. First and foremost is the recognition that the Web itself does not conform to our basic, semistructured data representation. As a consequence, we separate our discussion of challenges first into issues posed by semistructured data in general and then Web specific concerns.

6.3.1 Challenges attributing semistructured data

First, we note that in the formal model, attribution is modeled as external metadata set apart from data domains and relations on domains. Accordingly, in the algebra, we extended the data model by associating metadata with values and tuples (Sadri 1991) rather than incorporating attribution metadata explicitly into the relational schema (Dey, Barron, and Storey 1996; Dey and Sarkar 1996). In the semistructured context, it is easy to see how attribution could emerge as a metdata graph rooted to every node in the data graph through an "attribution label." Changes in query semantics as well as implications for a data representation that essentially duplicates paths in the graph need further thought.

Second, our intuitions about query composition and attribution composition, while analogous to their relational counterparts in the abstract, suffer from some difficulties in the details. In the formal model, the attribution for a composed query is defined by the attribution for the unfolded calculus expression. However, it is not clear that there is an equivalent for unifying two semistructured path expressions. Consider the case where one expression is a restriction and restructuring of the same graph (i.e. using the same nodes and labels).

A related problem, alluded to earlier, is the issue of recursive queries. Given a finite graph to begin with, we know that a path expression on a finite graph, recursive or otherwise, must return a finite set of nodes. However, the explicit paths that we can associate with the path expression for any given node, in the presence of a cycle, can be (countably) infinite. While there has been recent work on recursive queries in semistructured data (Abiteboul, Buneman, and Suciu 2000), finding a corresponding resolution for attribution will require some additional consideration.

Finally, graph reconstruction also poses a problem for our intuitions about granularity and aggregating attribution over subgraphs. Because variable assignment allows unrestricted (re)use of a given node (label) in structuring a result graph, there is no necessary dependency between the attribution of a node and the attribution of its children in a graph. For example, we could imagine constructing a result graph that associates hostel prices with hotel nodes. The attribution of the hotel node would have no bearing on the attribution of the hostel prices.

6.3.2 Challenges attributing the Web

Other challenges derive from the nature of the Web itself and the recognition that data on the Web today does not correspond neatly to any formal semistructured model. First, we know that the relational data model is value oriented. Every tuple instance is unique by definition. As noted earlier, in semistructured data, object identity causes value-oriented uniqueness to break down. On its own, this does not pose a problem, as the concept of identical values with different attribution appears in the formal model. A related problem that does emerge, however, is the question of duplicates. More generally, reflecting the Web's document-centric history, every node is represented (no weak duplicates), and order matters (Bray, Paoli, and Sperberg-McQueen 1997).[42] The need to reference order, both for querying and attributing, requires richer concepts.[43]

Apart from label order, the labels themselves pose some difficulty for being able to construct precise paths. Although standards for XML, in conjunction with XSL and Style Sheets, are evolving to address issues of meaningful, content-based labels, the Web today is dominated by HTML (Chamberlin et al. 2001b; Clark and DeRose 2001; Fernandez and Marsh 2001; Fernandez and Robie 2001; Grosso and Walsh 2000; Raggett 2000). Without special knowledge, then, there is a limit to the data that we can extract and attribute. Consider again hotelguide.com and their Web page on Tokyo hotels in Figure 6.2. While we hypothesized a semistructured representation in Figure 6.3, the data on the page really appears as the HTML that appears (in a slightly abbreviated form) in Example 6.10.

Example 6.10 HTML for hotelguide.com

A vision for the very near future of the Web calls for servers that return XML pages associated with style sheets to control presentation (Bray, Paoli, and Sperberg-McQueen 1997). Today, however, most sites, like hotelguide.com, still present HTML. Excerpted below is edited source from the page for Figure 6.2.

```
<body>
   <table width="100%" cellpadding="0" cellspacing="0" border="0">
   <tr>
      <font class="hotellist"><b>
      <a href="/html/searchengine/">
         ASAKUSA VIEW HOTEL
      </a></b>
```

[42] The Web (and HTML in particular) was originally conceived as a tool for sharing research literature. As a consequence, particular attention was directed towards formatting and presentation. So while an academic paper is, abstractly, composed of different sections, we might wish to ensure that "Section 6 Data analysis" comes after "Section 1 Introduction." Notice that our graph-based basic semistructured representation has no provisions for explicitly stating that one node or label is first in a sequence.

[43] The reader may note that the problem is not "duplicates" per say but rather one of "order." We merely use duplicates as an example of the need to define explicit order.

```
   </tr>
   <tr>
      <td width=32% valign="top">
         <font class="hotellist"><font size="1">
            3-17-1 NISHI-ASAKUSA, TAITOU-KU
         </font></font>
      </td>
      <td width=24%>
         <font class="hotellistsmall">
            single
            &nbsp;
            <a href="JavaScript: newWindow=currencyconverter">
               18000
            </a>&nbsp;
            JPY
         </font>
         <br>
         <font class="hotellistsmall">
            double
            &nbsp;
            <a href="JavaScript: newWindow=currencyconverter">
               20000
            </a>&nbsp;
            JPY
         </font>
      </td>
   </tr>
</table>
&nbsp;
... □
```

In HTML, the tags (labels) are structure-based rather than content-based. As a consequence, in writing a path to access particular items of data in HTML, we are forced to make certain assumptions about the order of fields as well as the data that we will find in those fields (Firat, Madnick, and Siegel 2000; Mendelzon, Mihaila, and Milo 1996).

We have continued to refer to www.hotelguide.com as a source, but in truth, the problem of identifying a source on the Web becomes much more complex than a relation name. URLs are clearly inadequate because of the temporal nature of data on the Web. Sites hosting dynamic content such as news or financial information are constantly changing. Even a URL with a path expression that specifies order may not suffice to concretely specify a distinct value or its associated attribution path. If we reference a granule by a path, does the path similarly name a source? In this case, a named source can contain a second source perhaps presenting a refined case of composition. (Buneman et al. 1997) has studied aspects of this problem in the context of keys for semistructured data, but the continual challenge will be to extend conclusions to the ad-hoc Web.

Other such pragmatic problems related to the ad-hoc nature of the Web and naming have to do with duplicate sites and whether replicas or mirrors are treated as distinct sources or the same source. Versioning and the temporal nature of the Web in general will also pose problems for attribution.

The Web today almost certainly foreshadows the future of data management. If nothing else, the metaphors carried from the print and publishing world onto the Web will continue in some form tomorrow. Meanwhile, as the Web continues to expand, incorporating ever more data, so to does the need for attribution as a mechanism for managing that growth, whether for search, intellectual property, or evaluating quality. For this reason, extending formal models of attribution (Cui, Widom, and Wiener 1997 (revised 1999); Motro 1996; Rosenthal and Sciore 1999; Sadri 1991; Wang and Madnick 1990) into the semistructured environment is the logical direction to look. The work by Buneman et al. is a terrific start (Buneman, Deutsch, and Tan 1998; Buneman, Khanna, and Tan 2000; 2001; Buneman, Tajima, and Tan 2001).

7 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.

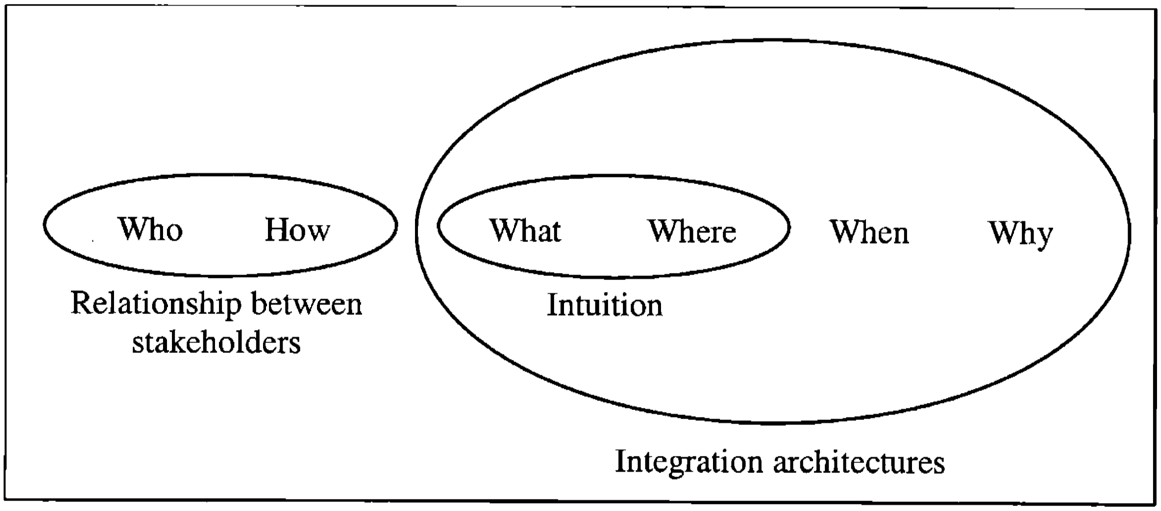


Figure 7.1 Attribution problem space (redux)

7.1 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.

7.2 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.

7.2.1 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." [44] 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).[45]

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

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

[45] 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.

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.

7.2.2 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 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)[46]. 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

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

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.

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.

7.2.3 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.[47]

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

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

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 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.

7.2.4 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.

7.2.5 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).

7.2.6 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.[48]

7.3 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.[49] 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.

7.3.1 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.

[48] 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).

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

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 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[50]. 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.[51] 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[52]. Shipping, not data, is FedEx's core business. To the degree that FedEx considers itself

[50] Consider firms such as Charles Schwab & Co., Inc. or Datek. Online Financial Services, LLC

[51] FedEx Corporation.

[52] Intershipper may be viewed at: www.intershipper.com.

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 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.

7.3.2 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.

7.3.2.1 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.

7.3.2.2 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.

7.3.2.3 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.[53]

[53] 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.

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.

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.

7.3.3 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.

7.4 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.

7.4.1 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).

7.4.2 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.

8 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).[54]

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.

[54] 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)."

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)."[55] 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.

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.

8.1 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

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

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.[56]

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 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?

8.2 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

[56] 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.

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").[57]

8.2.1 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 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.[58] 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.[59] 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.[60] 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.

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

[58] 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.

[59] 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.

[60] 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).

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.[61]

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 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[62] suggests that there is ample room for creativity at the physical level.[63]

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).

8.2.2 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.[64] 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).

[61] 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.

[62] Most commercial database vendors implement some version of the relational data model at the logical level.

[63] 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.

[64] The tongue-in-cheek reasoning argues that if database design is trivial, why do consultants get paid so much money to design them.

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).

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)."[65] 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).[66] 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.[67] Consider also the difference between the yellow pages and the white pages

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

[66] 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.

[67] 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

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.[68]

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 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.

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.

[68] 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.

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.[69]

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.[70] 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.[71]

8.2.3 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).

[69] 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).

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

[71] 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).

However, it is our contention that first, each index constitutes a distinct collection.[72] Second, alternative arrangements quickly become too exhaustive to print in a single publication for any database of significant size.[73] While a restaurant guide might provide an index that constitutes a different arrangement of the same selection, more complex collections such as travel guides often provide only abbreviated indexes that represent a more limited selection in an alternative arrangement.[74]

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.[75] 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.[76] Past technical limitations merely clouded the issue that eventually came to a head in *Feist* (Feist v. Rural 1991).

8.2.4 Electronic databases: disaggregation and appropriability

Modern information technologies disaggregate the product of gathering from the product of selecting and arranging.[77] There are many ways to arrange the same selection of data[78] and we might combine selected subsets from different collections to create a new whole.[79]

[72] 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.

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

[74] 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)

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

[76] 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).

[77] 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.

[78] 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).

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

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.[80]

Disaggregation is enabled through the power of abstract data definition and manipulation languages (Abiteboul, Hull, and Vianu 1995; Maier 1983; Ullman 1988).[81] In relational database systems, a single SQL (Structured Query Language) instruction both selects and restructures (arranges).[82] [83] 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.[84] 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).[85] 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.

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

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

[82] 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).

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

[84] 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).

[85] 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).

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.

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.

8.2.5 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.

8.3 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)

8.3.1 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.

8.3.1.1 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 8.1 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).

8.3.1.2 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).[86] 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.

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

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 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.

8.3.1.3 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.[87] 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).

8.3.2 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.

[87] 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.

8.3.2.1 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. 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.

8.3.2.2 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).

8.3.2.3 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[88] 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).

[88] 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.

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).

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.

8.3.2.4 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.

POLICY FORMULATION

> [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.

8.3.2.5 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.[89] 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 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).[90] 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

[89] See Section 3.1.2

[90] 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.

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 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.[91]

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).[92] 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.

8.3.3 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

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

[92] 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.

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.

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.

8.4 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.

8.4.1 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.

8.4.1.1 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).[93] 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

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

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 (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).

8.4.1.2 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.

8.4.2 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.

8.4.2.1 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).[94]

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.

[94] 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.

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.

8.4.2.2 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 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.[95] 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

[95] 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).

to license, at least in part, explicitly because of its interest in entering the market itself (Feist v. Rural 1991).[96]

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, 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.[97] Without creating the application first, the originality might never see the light of day.

8.4.3 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

[96] 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).

[97] 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.

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.

8.5 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.

8.5.1 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)

8.5.2 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.

8.5.3 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.

8.5.4 (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.

8.5.5 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).

POLICY FORMULATION

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.

8.6 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.

8.6.1 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.

8.6.2 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.

8.7 Objections

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

8.7.1 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.

8.7.2 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.

8.7.3 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).

8.7.4 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.

8.7.5 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).

8.7.6 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.

8.7.7 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.

8.7.8 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.

9 Conclusion

In this thesis, we explore technologies and 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 technology and policy perspectives, we divide 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 entire thesis. We then review our contributions and discuss both limitations and opportunities for future work.

9.1 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 technology or policy measures, we can better understand how the initiatives interact.

In Part one of this thesis, we focus on a technology-oriented approach to the questions of *what* and *where*. We first present a formal model of attribution that represents *what* as a query result and *where* as query inputs. Although our initial interest was sparked by data integration on the Web, we construct our model in terms of the well-understood, logical foundations of the domain relational calculus. Then, beginning with conjunctive queries, we define and evaluate properties of attribution for several different classes of queries. We consider conjunctive queries, conjunctive queries with θ-comparisons (excluding explicit equality), add explicit equality, add union, and finally add negation.

While the domain relational calculus offers a useful framework for developing our model, the definitions are not easily implemented. Consequently, we present an extended relational algebra for attribution. The extended algebra manages attribution in an inductive fashion. Metadata for specifying comprehensive, source, and relevant attribution is associated with every value in a relation; the metadata is updated and carried forward with every successive query operation. After showing some properties relating the extended algebra to the standard relational algebra, we verify that the attribution returned by the algebra corresponds to the attribution defined by the formal model for the same query.

Although our initial interest in attribution stems from the phenomenon of data integration that pervades the Web, we develop our model of attribution in the simpler but more well-understood framework of the relational data model. We conclude Part one of this thesis by returning to the semistructured data models that underlie the Web. We specify some general principles of semistructured data representation and manipulation and then discuss how our attribution intuitions might map onto this semistructured framework.

Part two of this thesis is directed at policy approaches to the management of the attribution problem space. 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.

9.2 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. Some of the prior work has focused on domain specific applications (e.g. geographic information systems (Lanter 1991; Woodruff and Stonebraker 1997)) and others have focused on general models. More recently, Buneman et al. (2001) has even developed a formal model for attribution in a semistructured framework. However, we feel that ours is the first to present the problem in a single framework, the dimensions of the attribution problem space, that articulates the relationship between different technology and policy approaches. In addition, we believe that this thesis does provide a number of contributions to both the existing technology literature and the existing policy literature.

The formal model defines three different attribution *types*. *Comprehensive* attribution refers to all query inputs. *Source* attribution refers to the specific inputs in which a specific value appears.[98] *Relevant* attribution asks which query inputs are used to define constraints or restriction conditions on a value of the query result.

We define several properties of attribution and provide a comprehensive analysis of these properties, covering each type of attribution for the full range of relationally complete expressions. We show that strict equivalence for source attribution breaks down under strict equality and that strict equivalence breaks down for all types of attribution upon introduction of union. Attribution composition is particularly useful because it demonstrates that attribution can be constructed inductively and carried forward with the query processing as well as drilled backwards in a step-wise fashion. We show that composition holds for all classes of queries through limited forms of negation and characterize those limited forms of negation.

Recognizing that we might wish to specify results or sources with varying degrees of precision, we introduce the notion of granularity to attribution. Granularity leverages the equivalence property of composition. Attribution is defined for a query result; result granularity attributes a specific subset of values in a result (*what* data is taken) by attributing a composed query that selects the desired values from the initial query result. Because a result granule can itself serve as a source for a composed query, we note the parallel concept of source granularity for specifying a subset of source values (*where* the data comes from).

While ours is not the first extended algebra to address attribution (Motro 1996; Sadri 1991; Wang and Madnick 1990), we prove a number of properties that are left unspoken in earlier work. Relating the extended algebra to the standard relational algebra, we prove that the extended algebra is closed and that it reduces to the standard algebra.

More significantly, our development of a formal model allows us to prove a number of properties not declared in other models. By defining attribution independently of the algebraic definition, we can show that the attribution algebra is a consistent extension of the standard algebra (Dey, Barron, and Storey 1996; Dey and Sarkar 1996). Finally, rather than defining the extended algebra and then simply stating that attribution is defined as whatever the algebra returns, we show that the algebraic attribution for a value in a query result corresponds to the formal model. The formal model allows us to express what is meant by attribution as well as some of its properties. The algebra provides a direct implementation of that model.

[98] For any given value (*what*) in a query result, the source attribution identifies the specific query inputs (*where*) from which the value in the result is drawn.

Independent of the technology analysis, 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.

9.3 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.

In terms of the formal model, there are a number of opportunities for further work within the existing model and for expanding the current model. In this thesis, granularity is mentioned only as an observation. We need to define granularity formally and define the relationship between attribution for the same query at different levels of granularity. Just as we speculate on converting between different granules, we might speculate on converting between different types of attribution. Finally, attribution refers to the substitutions for unique instances of values in tuples. Therefore, we should consider the role of functional dependencies.

We might also extend the model in at least two directions. First, we should consider whether the formal model can embrace a richer class of queries. The work on data lineage, for example, has extended to aggregation functions and more general classes of functions (Cui, Widom, and Wiener 1997 (revised 1999)). Second, we would like to consider parameterizing the model. Perhaps we could insert specific quality metrics or other measures that are a function of data attribution. (Rosenthal and Sciore 1999), for example, speak of the access constraints on integrated query results.

In considering the algebra, we first must find an algebraic definition for *relevant* variables that corresponds to the formal model. We found syntactic rules that captured a superset of the relevant variables, but had trouble defining a simple function that would support the formal definition.

We would also like to consider the eager, algebraic manipulation of attribution to manage aggregations or other more general classes of functions. Note that the Stanford work addresses the problem in a lazy manner. In addition, we could consider parameterizing the algebra to manage attribution-related metrics. Moreover, we might wish to explore whether the extended algebra is appropriate for managing other types of metadata such as that used experimental data collection (e.g. experimental apparatus, conditions under which the data was collected, etc.)

Extending the formal model to a semistructured data representation is also needed. As noted in Chapter 6, there are a host of considerations. Naming is a problem. As we commented earlier, how do we reference a source given that URLs are inadequate?[99] A second problem is the management of query composition and granularity. How do we frame these issues in an environment that allows graph restructuring?[100]

While the current thesis focuses on the theory, there is a great deal of opportunity in implementation. An initial algebraic prototype is described in (Lee, Bressan, and Madnick 1997). In the prototype, attribution is calculated in an eager manner and carried forward with every value. We would also like to implement the algorithm for attribution composition and explore attribution composition as a hybrid lazy-eager approach. Only one step of the attribution is calculated and propagated while enabling a step-wise backwards trace. In the context of the Web, we might consider an attribution Web service to support attribution tracing between integrated query results.

As with the model and algebra, we could extend the policy analysis in a number of ways. 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.

[99] In Chapter 6, we observed that the temporal nature of data on the Web as well as dynamic Web sites and personalization (e.g. Web site as modified based upon cookies) can all affect the content referenced by a URL).

[100] As noted in Chapter 6, the formal model leverages the value-oriented characteristic of the relational data model. In semistructured data, however, different paths (i.e. different structure) can return the same values from the same domains.

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.

REFERENCES

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.

References

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.

References

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).

REFERENCES

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.

REFERENCES

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.

References

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.