March 9, 2011

Transparent and Commercialized?: Managing the Public-Private Model for Data Production and Use

Shubha Ghosh, University of Wisconsin - Madison

Available at: https://works.bepress.com/shubha_ghosh/6/
TRANSPARENT AND COMMERCIALIZED?:
Managing the Public-Private Model for Data Production and Use

Shubha Ghosh
Vilas Research Professor of Law
University of Wisconsin Law School

I. Introduction

While it is common to make reference to the “information age,” a more accurate description of the commercial, political, and social context we live in is the “age of data.” Governments, companies, and individuals generate data at such a pace that the management of data creates quandaries. At the same time that we see the confidentiality of data, political and economic forces push to make data open, transparent, and exploitable.

This Article explores the contours of the forces and counterforces that surround the commercialization of data and presents a model for regulating and constraining the policy questions raised by data commercialization. The specific focus of this Article is on the connection between data collected and disseminated by the government and that generated by private transactions, whether commercial or in social settings such as online interactions.


2 See, e.g. Barbara J. Evans, Authority of the Food and Drug Administration to Require Data Access and Control Use Rights in the Sentinel Data Network, 65 Food and Drug Law Journal 67, 73-74 (2010)(analyzing FDA database on pharmaceutical usage). See, also, examples discussed at notes 8-16 infra. A compelling illustration of this
public-private model for data production and use and propose a contracting model, based on open source licensing, to govern the dissemination of data.

The Year 2010 tested the limits of the transparency of governance, access to information, and the governance of data. Wikileaks raised issues of how transparent open government could and should be.³ But this controversy overshadowed more pervasive concerns about the commercialization of data. Throughout 2010, social network sites challenged the balance between private and public information.⁴ Highlighting these concerns was a memo from Google, leaked in August, 2010, which revealed many of the contemporary dilemmas over the commercialization of data.⁵ The memo described various business strategies Google had considered that utilized information on consumption patterns obtained from its users of email and search engine services.⁶ The company, a trend setter in the development of markets based on digital technologies and information services, expressed hesitation in compromising privacy for profit while describing innovative platforms for the exchange of data and advertising in order to target specific needs of the company’s customer base. The self-described “vision
statement” portends the future of consumer generated and business processed data as the future source of commercial value in information based economies.\(^7\)

The issues raised by Google’s vision of the commercialization of data goes beyond accepted Internet transactions, such as emailing, searching, and social networking. Innovations in the processing of genetic information,\(^8\) particularly as applied to personalized medicine and the possibility of genetic profiling, raises analogous questions of the extent to which commercialization of data is constrained by social concerns over privacy and individual autonomy and legal concerns over the scope of property rights. The broad question raised by both Internet transactions and genetic information is to what extent can information generated by individual persons be the basis for business models for the collection, aggregation, and processing of information.\(^9\) This question, however, extends beyond the realm of new information technologies. For example, an issue in the 2010 Census was the recording of same sex couples as married on the census form.\(^10\) The recognition of marital status marks a progressive turn in equalizing the treatment of same sex and different sex relationships. At the same time, advertisers heralded the shift because it allowed for more effective marketing and targeting of advertising dollars to same sex couples.\(^11\) The Google memo, personalized medicine, and the Census example, arising from very different technological, legal, and social environments, raise a common question: how can data and information be used in a liberal, democratic society based on a market economy? Is it possible to reconcile values of transparency and openness with those of commercial gain and ownership? This paper addresses these questions as matters of legal entitlements and institutions as well as one about the construction of social meaning.

My thesis is that contemporary debates about the commercialization of data, illustrated by the examples above, reflect an inherent tension between democratic values of transparency and accountability and market goals of wealth creation. On

---

\(^7\) Id.


\(^9\) See “A Sea of Sensors,” The Economist, November 6, 2010, at S6-S7 (describing information gathering sensors and their relevance for commercialization, pricing, and marketing); “London: Turning Access Into Apps,” Time, January 17, 2011, at 6 (how state collected information serves as the basis for iPhone applications).


the one hand, democratic society demands transparency and the disinfecting influence of sunshine. This demand mandates that information and data be open and accessible. On the other hand, wealth creation through markets requires that data and information be manipulated and transformed, often in ways that distort underlying values of openness and the meanings constructed around individual quanta of data and information. This article explores this tension, demonstrating different ways in which it is identified and reconciled and concluding that some form of open source licensing, with trumps for the protection of personal privacy and autonomy, would be the appropriate business model through which data and information are commercialized.

The previous paragraph deliberately conflated the terms data and information in usage. This conflation reflects ordinary language usage and illustrates some of the tensions in coherently reconciling the conflicting values over commercializing data. In fact, drawing a distinction between data and information may serve as a foundational step in understanding the proper arrangement for commercializing data.

For example, data.gov is a web site created by the Obama Administration which makes accessible a range of government databases, covering such topics as atmospheric data, employment statistics, and government expenditures. The terms of use for the site do not prohibit the distribution of data or the commercialization of the data, as long as the data is distributed without restrictions on reuse of data. Furthermore, the terms make a distinction between data and information. Data is defined as “values or sets of values representing a specific concept or concepts.” According to the site, data becomes information when combined to extract meaning and put data in context. Put another way, data is raw or uncooked information while information is processed or interpreted data. This distinction between information and data is important in light of the prohibitions on restrictions on reuse of data. Implicitly, the terms of use would not prohibit restrictions on reuse of information. By making a distinction between data and information, the government leaves a domain that is open and accessible (data) while allowing the possibility for proprietary domain where market transactions can be purely private and value can be appropriated (information).

The distinction between data and information is just one way through which the conflicting values of privacy, autonomy, and proprietary rights can be

reconciled. As I demonstrate in greater detail in Section Two of this Article, other distinctions, such as between original and unoriginal works or between information merchants and non-experts, can serve to construct a market for data. At the heart of these distinctions are three normative positions about how to reconcile data transparency with data commercialization. I will refer to these three respectively as: (i) the classic liberal perspective (referred to in this Article as “the liberal perspective”),

14 (ii) the autonomy perspective;

15 and (iii) the fairness perspective.

16 Stated briefly, the liberal perspective would conclude that there is little tension between data transparency and data commercialization because a fully autonomous person is one who acts in both the political and market arenas. The autonomy perspective, however, would see a potential conflict as commercialization of data leads to a confounding of personal and public spaces that can distort meanings and identity. Finally, the fairness perspective would find that data transparency and data commercialization are irreconcilable since market commercialization will lead a corporate culture of commodification to trump other more human and socially desirable values. These three positions are ideal categories, useful in sorting through the various arguments identified in this Article. Each perspective casts some light on the policy issues raised by the commercialization of data and serves in shaping how data commercialization can be organized in an effective and desirable manner.

The structure of this Article is as follows. Section Two presents the economic issues raised by the commercialization of data, developing the various normative points made in this Introduction. Section Three builds on the economic realities by turning to the legal structure of data markets, rooted in intellectual property and other information-regulating laws, such as privacy and security. Section Four presents the argument for an open source licensing model for data commercialization. Section Five summarizes and concludes.

II. The Economics of A Data Market

14 For an example of what I mean by the classic liberal perspective, see Archon Fung, Mary Graham, & David Weil, Full Disclosure: The Perils and Promise of Transparency 171-173 (2007)(describing ways in which government policy can effectively empower people as market and political actors); Alfred D. Chandler & James W. Cortada, The Information Age: Continuities and Differences, in A Nation Transformed, supra note 1 at 298-299 (describing the development of information and digital technologies as an American story, celebrating individual liberty).

15 For an example of what I mean by the autonomy perspective, see Daniel J. Solove, The Digital Person: Technology and Privacy in the Information Age 8-9 (2004)(describing the problem of information privacy as a matter of individual loss of control, especially in face of a large bureaucracy).

16 For an example of what I mean by the fairness perspective, see Edith Brown Weiss, In Fairness to Future Generations 258-259 (1989)(advocating intergenerational equity as a guiding principle in the management of knowledge systems, which include archiving and preservation of data about natural and cultural worlds).
Theoretical analyses of markets for data suggest several foundational problems for the commercialization of data. The problem rests on turning data into an identifiable and quantifiable commodity that can be priced and rationed, like other commodities traded in a market system.\(^{17}\) The theoretical argument against a market for data rests on the economic assumption of perfect information.\(^{18}\) One assumption for a market to operate efficiently is the existence of perfect information among actors in the marketplace. If information is asymmetrically known or is missing, then markets may break down.\(^ {19}\) The problem is how information itself is generated and distributed in order to ensure well-functioning markets. In order for there to be proper incentives for information to be acquired, information has to be to a certain extent proprietary. But if information is proprietary, it is no longer perfect. This paradox\(^ {20}\) can be resolved through an appropriately designed legal regime to govern the accumulation and distribution of information. The debate among scholars over how the legal regime should be designed is the focus of Section Three.

This well-known argument about the conundrum of producing perfect information in the marketplace becomes more nuanced when data and information are distinguished. Data is, according to the definition from data.gov and other sources, more rudimentary than information, which is understood as the interpretation of data. The assumption of perfect information in markets is about information, and not data. But once the relationship between information and data is understood, the assumption of perfect information is essentially two assumptions, one about the distribution of relevant data about the marketplace and the other about the relevant distribution of interpretation of primitive data. What was described in the previous paragraph as the question of the design of the appropriate legal regime for regulation of information is about two related issues:

\(^{17}\) For an analysis of the purpose of commodification and its limits, see Margaret Jane Radin, Contested Commodities 79-102 (2001). Commodification of data is central to the argument in favor of price discrimination which holds that price for goods and services should be tailored to individual preferences and willingness to pay. See, e.g., John P. Conley & Christopher Yoo, Nonrivalry and Price Discrimination in Copyright Economics, 157 U. Pa. L. Rev. 1801, 1814 (2009)(describing role of price discrimination in markets for information, with particular reference to genomic information).


\(^{19}\) See Akerlof, supra note 17 at 491.

\(^{20}\) For an analysis of this paradox, see Sanford J. Grossman & Joseph E. Stiglitz, On the Impossibility of Informationally Efficient Markets, 70 Amer. Econ. Rev. 394-408 (1980)(showing how incentives to obtain perfect information are weakened in perfect markets).
designing institutions for the regulation of data and those for the interpretation of data when processed as information.\textsuperscript{21}

In this section, I explore the assumption of perfect information in competitive markets in order to explicate the underlying economics of a market for data. This section begins with a summary and critique of two well-known theories of legal institutions governing information: the authorship theory of Professor James Boyle\textsuperscript{22} and the notion of lex informatica, developed by Professor Joel Reidenberg.\textsuperscript{23} I show that each fails to adequately address the relationship between data and information in the process of commercialization. I propose that the gaps in these two approaches can be filled by a deeper appreciation of what I call “the life of data,” or the set of practices that support the commercialization of data from its creation to its aggregation and processing into information. The section concludes with a description of the economics of data commercialization that takes into account the “life of data.” I motivate this economic analysis through the character of Sherlock Holmes,\textsuperscript{24} a literary model for how contemporary data commercialization operates.

A. Perfect Information, Authorship, and Information Law Merchants

Economists Sanford J. Grossman and Joseph Stiglitz are credited with first describing the paradox of perfect information in market systems.\textsuperscript{25} If a market is to be efficient, information has to be perfect in the sense that actors know the relevant parameters of the economic environment, such as tastes, costs, prices and incomes, in order to make relevant decisions. If information is imperfect, perhaps because it is asymmetrically distributed with some individuals being more informed than others, then market choices may be distorted and exchanges that increase social wealth may not occur. The problem is that since market systems rest on private incentives, the problem arises of how information is acquired and generated in the

\begin{footnotesize}
\begin{enumerate}
\item The distinction is one between data gathering and data aggregation and processing into information. See R.L. Ackoff, From Data to Wisdom, 16 Journal of Applied Systems Analysis 3-9 (1989)(drawing a hierarchy from data to information to knowledge to understanding to wisdom).
\item The character of Sherlock Holmes was created by Sir Arthur Conan Doyle in 1887 in the book A Study in Scarlet. The reference here is inspired by economist Edward Leamer, who wrote that empirical economists (and by extension social scientists) should follow the model of Sherlock Holmes and neither fit the data to the theory nor ignore theory and focus solely on facts. Instead, the social scientist should move back and forth between theory and facts. See Edward Leamer, Let’s Take the Con Out of Econometrics, 73(1) Amer. Econ. Rev. 31-43 (1983). I am suggesting that Sherlock Holmes is the model also for the data miners in the contemporary world.
\item See supra note 20.
\end{enumerate}
\end{footnotesize}
first instance. Information has to be proprietary, to a certain extent, in order for it to be effectively collected. Put bluntly, the market for information requires perfect information in order to function, but the returns to acquiring information are driven to zero if markets are operating perfectly. Hence, the circularity or paradox of perfect information.

This circularity explains why markets should not be expected to function perfectly and why institutions, such as government regulation, will be needed to correct inevitable market failures. Furthermore, institutions that supplement the market will be necessary to ensure that information will be generated and collected. These institutions might include banks, brokerages, regulatory agencies, and other information aggregators that engage in the creation, aggregation, assessment and dissemination of information. This section explores two theoretical approaches for understanding these institutions for the generation of information, the first associated with Professor James Boyle and the second with Professor Joel Reidenberg. I present and analyze each and then suggest an alternative approach that takes into consideration the distinction between data and information and their practical uses, what I call the “life of data.”

1. Professor Boyle and Authorship

Professor Boyle builds on the paradox identified with Grossman and Stiglitz to develop a theory of institutions pertaining to ownership over information. The paradox of information in market systems is resolved in part by identifying certain uses of information as correct and appropriate and others as undesirable and in need of deterrence.26 For example, prohibitions on insider trading restrict who can profit from certain types of information and when. The criminalization of blackmail, to take another well worn example, limits incentives to discover private secrets in order solely to profit from their nondisclosure.27 Indirectly, the criminalization of blackmail also leads to transparency while protecting a realm of privacy. Finally, intellectual property law, particularly copyright, protects certain types of information, namely that found to be original, novel, or in other ways innovative.28 Through these different legal regimes, Professor Boyle concludes, property rights in information are created for the purpose of creating markets that function effectively.

26 See supra note 20 at 397.
27 See Boyle, supra note 22 at 235 (insider trading); supra note 22 at 69-70 (blackmail).
The definition of property rights over information, as described by Professor Boyle, entails the construction of authorship and the authoritative use and creation of information. Property rights create a dichotomy between recognized and unrecognized forms of transaction, with insider trading, blackmail, and unauthorized copying falling into the latter category. Digging deeper into this dichotomy, claims of transparency and open access assume some realm in which information is proprietary and undisclosed. As Professor Boyle points out, the boundary line between authorized and unauthorized is an unpredictable one in practice, resting on normative assumptions about efficiency and distribution that are not examined. Restrictions on insider trading, for instance, are justified as leveling the playing field in financial transactions. In reality, however, the field is never level with some purchasers necessarily being more sophisticated and informed than others. Property rights over information rest on shaky and indiscernible normative premises.

The problem over authorship can be seen in the context of intellectual property over facts. The United States Supreme Court famously stated that facts are not protected by copyright in the 1991 Feist decision, but lower courts quickly began to make distinctions between facts that cannot be protected and facts that could. Leaving facts open and in the public domain was justified on the need for transparency and the use of facts as building blocks. Facts were deemed not to be created and therefore not subject to ownership as intellectual property. But courts recognized that some information deemed as factual was in fact created (consider a statistical index which is a composite of more basic data such as sports scores). The notion of authorship would justify and necessitate the recognition of property rights in created facts. The treatment of facts as intellectual property illustrates the malleability of the distinction between authoritative and non-authoritative, and hence between what information is open and what is closed.

The slipperiness of the concepts of authorship and property rights is reflected in the notion of information itself. Professor Boyle (and before him Grossman and Stiglitz) wrote about information, but as we have seen, there is a distinction between data and information. For example, the web site data.gov makes such a distinction and contemporary policy debates speak of the regulation

29 See Boyle, supra note 22 at 18-19.
30 See Boyle, supra note 22 at 82.
32 See, e.g. CDN, Inc. v. Kapes, 197 F.3d 1256 (9th Cir. 1999)(recognizing copyright in price data).
33 499 U.S. at 346.
34 See, e.g, Kregos v. Associated Press, 3 F.3d 656 (2nd Cir. 1993)(finding that tabulation of sports statistics was protected by copyright).
of data. The distinction between data and information is one between the raw and the processed. Data is defined as concepts that represent values. Information is the interpretation of this data through inference or other analytical methods. Data presumably is accessible to everyone while information is not. Once again, the dichotomy represents a distinction between a realm of openness and accessibility and a realm that is proprietary and subject to property rights. This dichotomy is necessary in order for markets to operate, but is one that is ultimately arbitrary.

Professor Boyle’s analysis is helpful in identifying the role that authorship plays in the definition of property rights. But given that this definition is ultimately arbitrary, there is perhaps very little that the theory tells us about the commercialization of data. There are two problems.

First, Grossman and Stiglitz’s argument, the starting point for Professor Boyle’s, is that effective markets for commodities requires perfect information. The problem, however, is assessing the market for information, which arises when property rights are defined. Is this market purely a secondary one, whose effectiveness is gauged by how well it facilitates the market for goods and services, the primary market? Or is there some way to assess the market for information itself as a primary market? The Google memo illustrates this dilemma as its authors attempt to gauge what data can and cannot be commercialized and how. The description of the memo as a “vision statement” may indicate how mercurial and evasive the answer to these questions may be. The problem, once again, is the arbitrariness with which property rights are defined.

Second, commercializing data assumes that data can be turned into a commodity for exchange. A commodity has to be standardized and measured in order for exchange to be meaningful. If parties to an exchange have different conceptions of what the underlying commodity is, common sense and legal doctrine would cast a pale on the legitimacy of the exchange. But as indicated above, what constitutes data as distinct from information, which is processed, may itself be arbitrary, a reflection of the bluntness of authorship in defining property rights. The Google memo, according to this argument, is not so much visionary, but misguided. Commercialization of data would, therefore, be like mining for fool’s gold, any value would ultimately be illusionary.

These conclusions undermine the practice of commercializing data. Advertisers view data as valuable to their venture of targeting products and
services and more individualized pricing. The value seen by advertisers in turn provides the foundation for various business models for selling goods and services (think Amazon.com) or information itself (think Facebook.com). Furthermore, data has value in better assessment of risks (think safety or accident data), the detection of fraud (think accounting data), or the assessment of professional performance (think test data or data on government responsiveness and accountability to social and environmental problems). Commercialization of data attempts to tap this value by creating markets for data and for information, data in its more processed form. While it is true that once property rights are defined, they may have an arbitrary quality and justification that can be malleable and manipulated. But the organization created by the definition of property rights supports institutional arrangements that make the use of data, including its commercialization, possible. Authorship has its limits as a concept for defining property rights, but its limitations do not mean that commercialization of data is impossible or incoherent. Rather, we need to look to other justifications to understand how commercialization of data can take place. In the next section, I examine the concept of lex informatica as one alternative. While this concept is also limiting, it will be useful in understanding the third approach I examine, that of the “life of data.”

2. The Information Law Merchant Confronts Data

Professor Joel Reidenberg introduced the concept of lex informatica in an article about information policy in the 1990’s. Lex informatica is a play on the term “lex mercatoria” or the law merchant, a reference to the informal creation of contract rules by merchants the maritime trade of the Fifteenth and Sixteenth Centuries. The law merchant created established contractual and exchange through repeated interactions among players engaged in standardized transactions. Professor Reidenberg analogized that the Internet was also creating an informal set of rules through repeated interactions among players through websites. The primary mechanisms for the creation of lex informatica are terms of use and practices by gatekeepers such as bloggers and web hosts. The resulting lex informatica reflects accepted practice in using data and information contained on the Internet.

37 See Reidenberg, supra note 23 at 554.
38 See id. at 561-562.
Since the concept of lex mercatoria is an evolutionary one, it is impossible within our current time frame to assess the importance and the effectiveness of lex informatica. But the concept is an important one for understanding how the rules and regularity of legal institutions are created in order to facilitate the commercialization of data. The appeal of the concept is that it does not, at least ostensibly, rest on an abstraction like authorship as a baseline for defining property rights. Instead, lex informatica looks to actual practice to see how regularity and norms are created in the use and exchange of data and information. This move is an important one for the purposes of this Article.

Lex informatica, however, does have one feature in common with the notion of authorship. Both assume a central figure or set of figures which shape how the legal rules are formed and understood. The notion of authorship assumes an archetypal author whose originality is a template for understanding what information is protected. Similarly, lex informatica assumes the figure of the merchant, who traverses the Internet engaging in transactions and thereby defining norms and rules. As with the author, there is a certain arbitrariness with the figure of the merchant with the resulting malleability and loss of attention to the normative premises underlying the rule. To be fair, this malleability might be the attractive feature of lex informatica as it suggests a more evolutionary development of property rights. At the same time, lex informatica may simply be another manifestation of the author, here not so much as creative force, but a transactional one.

A very different criticism of lex informatica is not the abstraction of the Internet merchant, but its clear identifiability. Professor Reidenberg sees terms of use as one of the key factors in shaping lex informatica, and terms of use are the product of attorneys working for communications companies, such as Time-Warner or Google. Lex informatica is predictably a product of what benefits corporate interests as opposed to those of other actors in the Internet. This predictability translates into contractual terms that limit the rights of users, allow the archiving and collection of data from users, and facilitate one sided transactions over the exchange of data. Corporate gate keepers can know you, but you do not necessarily know them or what they do. This picture is needless to say too one-sided and ignores democratic impulses that may limit a purely corporate form of lex informatica. Nonetheless, the concept of lex informatica, as distinct from that

---

40 See Reidenberg, supra note 23 at 564.
41 See Reidenberg, supra note 23 at 570.
of authorship, points us to the real politik of how commercialization of data is organized and shaped.

Google’s vision statement evinces the processes underlying lex informatica. Read as blueprint, the statement illustrates how practices are determined and shaped. Arguably, the statement dismisses some pro-consumer approaches to deal with data commercialization (such as allowing users to opt out from advertising for a fee). But it would be wrong to describe the statement as reflecting purely corporate interests. Instead, the document suggests a deliberative process with Google addressing the concerns of constituencies both internal and external to the company. When understood this way, the Google memo represents a richer form of lex informatica with the merchants taking into consideration the concerns of customers and users.

From this perspective, the distinction between data and information made by the website data.gov exemplifies a compromise between public accessibility and commercialization. The distinction is not based on an artificial and malleable concept of authorship but on a practical distinction between inputs, such as data, and outputs, such as information. The terms of use of data.gov do not limit commercialization of data, but do prohibit contractual or other mechanisms that would limit access to data. These parameters mean that data is never proprietary. On the other hand, information is not restricted since it represents the interpretation and processing by the user. To the extent this distinction becomes part of lex informatica, it represents a norm that allows for transparency of data as unprocessed information but allows for information to provide a source of commercial value through the refinement of data. In this way, the commercialization of data can occur at various levels and through different types of legal structures and market institutions, one for the commercialization of data as raw input and one for information as processed data.

What lex informatica draws out attention to are the practices over the marketing and use of data. In the next section, I build on this aspect of lex informatica to focus on the “life of data.”

3. The Life of Data

The paradox identified by Grossman and Stiglitz requires a resolution through the design of legal institutions that govern information. The approaches of authorship and of lex informatica have limitations in resolving the paradox. Part of the problem is drawing distinctions between data and information, distinctions that
reflect considerations of authorship, of norms of transparency, and of pragmatism. What is missing in the discussion is a good sense of the “life of data,” the actual ways in which data is generated and used. This topic is the focus of the rest of Section Two.

Information is not uniform and different types of information may be protected by different types of regimes.\(^4\) For example, in a market context, information about price and product quality might be viewed as open and transparent while information about a firm’s technology or a consumer’s preferences may be proprietary. The scholarly literature on the legal regulation of information has moved in part to recognizing the granularity and diversity of information. Accompanying this move is a recognition of the rich institutional environment in which information is produced and consumed. Markets for information, like all markets, exist in a rich ecosystem, and the dynamics of these institutions must be understood to appreciate the manner in which information is produced and disseminated. Joel Reidenberg’s work on lex informatica, by analogy to lex mercatoria, is one example of this deeper understanding. Under lex informatica, technology makers and keepers determine social norms that become a model for legal regulation of information. Reidenberg’s work turns to technology as the basis by which regulation occurs.

This paper also examines a type of lex informatica, but does not rest its formation solely on technologists. Like lex mercatoria, lex informatica develops from market and social contexts in which rules and conventions, of various degrees of formality, generate governing laws. In the case of lex mercatoria, these governing laws regulate commercial transactions. In the case of lex informatica, the governing laws modulate the manner in which information is generated, disseminated, and commercialized.\(^4\) My premise, however, is that this is not just a matter for the technologists. As the Census 2010 and data.gov examples show, this is a matter of technology embedded in broader social, political, and economic contexts. The next subsection draws a sketch of these contexts.

B. The Mechanics of Commercializing Data

1. The Model of Sherlock Holmes

---

\(^4\) See Spence, supra note 28 at 295-298 (describing range of legal protections for different types of information).

\(^4\) See Reidenberg, supra note 23 at 581.
The model I have in mind for lex informatica is not the contemporary technology manager but Sherlock Holmes. As a detective, Mr. Holmes was a processor of data and generator of information that yielded answers to seemingly intractable questions. I will not provide any spoilers here about his sleuthing, but most readers are perhaps aware of how the detective was able to discern whole life histories from simple facts like a scar or the way in which a person entered a room. Dr. Joseph Bell, the professor of Holmes’ creator Arthur Conan Doyle, was legendary in being able to make diagnoses from the minutest of details. Holmes is a data producer and a data consumer. His detection involved gathering and creating data as inputs into his deductive method and also as a passive consumer collecting information about the world. As a commercializer of data, Holmes made a living, of sorts, through the sale of his services. There are some records of monographs he published on cigar ashes and other marginalia. But his primary metier was using his craft at parsing data to solve puzzles and capture criminals.

The contemporary model of the commercializer of data starts with Sherlock Holmes. But contemporary data entrepreneurs go beyond the mere sale of services (although there is a rich market for data forensics). Data entrepreneurs package data through websites and other media so that users of all sorts can generate more data in a seemingly exponential growth. The result is a counter-Malthusian economy where the geometric growth of population is matched by the exponential growth of data and information.

Sherlock Holmes’s contemporary environment is captured in the postmodern literary genre of science fiction. The so-called literature of information, including such prominent novelists as Thomas Pynchon and William Gass, but also popular figures like Philip K. Dick, reveal the complexities of this counter-Malthusian economy. “Information overload” is the popular buzzword, but at a more sophisticated level the question is how the generation and consumption of

46 Id. at 55-56.
48 See, e.g., Gail M. Cookson & Carole Longendyke, Data Forensics, 29 Md. B.J. 66, 66 (Jan./Feb. 2006)
49 Thomas Malthus, describing agrarian economies, posited that human populations would grow exponentially while food supply would growth arithmetically, causing downward pressures on human populations. The counter Malthusian view described here suggests that the growth in data about individuals will grow faster than the population. See discussion in note 1 as illustration of the data deluge.
50 See Joseph A. Conte, Design and Debris: A Chaotics of Postmodern American Fiction 193-194 (2002)(describing how some contemporary American authors respond to the superabundance of information).
data affects the mindset. Sherlock Holmes was a methodical figure, a marriage of Cartesian rationality with British empiricism. In the counter-Malthusian economy of data, method and madness run in tandem as “supercrunchers”\(^{51}\) generate the next generation of results. Is everything connected or is all data the result of random processes? This is the question posed by novelists like Thomas Pynchon.\(^{52}\)

The lived experience is one where anything goes, all data is there to be used, processed, regressed, and parsed into a mix of innuendo, rumor, hypotheses rejected, statistical significance gauged and reported.

2. Creation, Aggregation, and Dissemination

A picture of the counter-Malthusian economy emerges from asking questions about production, consumption, and ownership. At the outset, the line between production and consumption with respect to information is a blurry and perhaps non-existant one, as many scholars have demonstrated.\(^{53}\) If data is just an input to produce information or more data, than a consumer of data is arguably just another producer in a never ending value chain. Nonetheless, I use the labels producer and consumer here to distinguish between the questions “where do data come from?” and “where do data go?” These questions are useful in providing some order in understanding the process of commercializing data.

1. Production. As Sherlock Holmes demonstrates, anything, from the smallest fiber of hair to an innocent movement of the eyes, can be a source of data. But data can come in many varieties and forms. As the data.gov site illustrates, data processed can generate information, which captures the inferences drawn as data is moved through a scheme of review and analysis. Furthermore, data and information drawn from data may be fixed in various ways that can generally be referred to as records which serves as storage and packaging of data and information. In addition, data and information can be arranged in various ways, such as through writings and publications that report on the analyst’s findings or in databases which serve to package data in ways that can be more readily parsed and analyzed for the generation of more information. Data production can best be divided into raw data and cooked data. The line between raw and cooked is one determined by normative decisions about how data should be regulated, as

\(^{51}\) Ian Ayres, Supercrunchers 244-245 (2007) (data access and analysis as democratic).


\(^{53}\) See, e.g., James G. Webster, User Information Regimes: How Social Media Shape Patterns of Consumption, 104 Northwestern L. Rev. 593, 608-610 (2010) (demonstrating how users of information both consume and shape social media); Joseph P. Liu, Copyright Law’s Theory of the Consumer, 44 B.C.L.Rev. 397, 411 (2003) (showing that copyright law views consumers as either passive users or as active creators).
discussed in greater detail under ownership, below. But framing the question in terms of raw versus cooked aids in understanding how normative questions of ownership and regulation are determined.

2. Consumption. As pointed out before, consumers of data are often intermediaries who are in turns producers of the next generation of data and information. In some instances, however, consumers of data may be pursuing data for its own sake. A consumer may simply want to know the answer to a question, such as one’s ancestry or presence of a disease in one’s body or something more mundane like the hours for a restaurant or museum. In terms of commercializing data, consumers may often be the construct of the marketplace as data entrepreneurs define a niche within which the activity of data retrieval and usage occurs. Bulletin boards, blogs, discussion groups are all examples of how consumption can be constructed by the data entrepreneur in order to generate another level of data on which value can be created. As data.gov illustrates, simply providing access to data may create consumption although with the case of government data, such demand may arise from political and economic needs to understand one’s rights or how one’s tax dollars are being spent.

3. Ownership. Data commercialization rests on identifiable legal rights and duties among parties in data related transactions. The challenge is identifying the structure of these rights and duties. In the Sherlock Holmes example, the entrepreneurial detective pursued a business model based on the sale of his detection services partly because securing property rights in the inputs of his business was impossible. The clues on which he built were in the public, in plain sight; his craft was in identifying and analyzing them. Contemporary data entrepreneurs need to negotiate a panoply of legal regimes in order to negotiate data-related transactions. Privacy and security restrictions may protect personal and other sensitive types of data. Trade secret law protects commercially valuable information within a firm. While data is unprotected under copyright law, the original arrangement, selection, and coordination of databases can be protected by copyright. Furthermore, contract law might offer some protection to data on a

---

54 See Section III, infra. The descriptive metaphor of raw and cooked is taken from anthropologist Claude Levi-Stauss, who made a distinction between the raw, or natural, and cooked, or cultural, in social systems. In this context, data is the primitive, or raw, and information is the cooked. The process of going from raw to cooked occurs through the processes of commercialization that are identified in this Article. For a related use of these terms, see Peter Lee, The Evolution of Intellectual Infrastructure, 83 Wash. L. Rev. 39, 41n16 (2008)(using the distinction between raw and cooked to distinguish between shared and proprietary resources).

55 Holmes’ capture of the value of his processing of information through providing detective services parallels the use of services to capture value in open source software. See Niva Elkin-Koren, Tailoring Copyright to Social Production, 12 Theor. Enq. L. 309, 321 (2011)(describing open source business model to obtain revenues through distribution of services).
bilateral basis. Finally, patent law might offer some protection to the methods of how data are processed and accessed although process patents are currently under scrutiny and may be subject to limitation. In short, the ownership structure of data is a complex one. Proprietary in some dimensions and legally unfettered in others, data ownership reflects the contradictory roles of information in transactions as both the basis for open and liberal markets and an investment that is expected to generate a return.

With this theoretical analysis of the counter-Malthusian world of data and information, we turn in Section Three to understanding the legal structure that governs the commercialization of data.

III. The Legal Structure of Data Exchange

The legal structure governing data affects how data can be created, distributed, and consumed. As the discussion so far indicates, under most legal systems, data is neither completely accessible nor completely proprietary. Instead, as this section demonstrates, the governance of data has many dimensions which attempt to balance the values of access with those of stewardship and management. This section does not present the ideal legal structure for data protection. The goal is to describe the legal regime as it exists in the United States with some comparison from the European Union, which has a decade old directive on database protection. This description of the legal regime will be assessed with a normative framework for addressing transparency and markets. The descriptive and normative frameworks will also structure the relationship between data and information, an important distinction in the practice of data commercialization.

An account of the legal structure of data can be divided into two sets of issues, ownership and data transactions. Intellectual property law grounds the ownership issues, particularly copyright, trade secret, and patent. The structure of ownership determines who can license and transfer data and what types of data can be subject to commercialization. Privacy law and norms as well as those of security are legal structures that govern permissible data transactions, such as the selling of personal data to advertisers or the collection of data for the purposes of profiling. State creation and ownership of data is an important piece of the legal structure with data.gov being just one example of the role of the government in creating, disseminating, and consuming data. The role of the state will be considered in the discussion of ownership, specifically in the analysis of Open Records Acts, which introduce another layer to the relationship between data and
information. As I will show in this section and in the case examples of Section Four, the government is a key player in such situations as the commercialization of real estate data and census data.

Understanding the legal structure is a step towards assessing data commercialization against a normative framework. As explained in Section One of this Article, there are three normative perspectives one can take towards the practices of data commercialization:

(1) the liberal perspective;
(2) the autonomy perspective;
(3) the fairness perspective.

Each perspective addresses the twin issues of transparency and commercialization in the management and use of data.

Under the liberal perspective, transparency and commercialization are not incompatible because the market realm provides a form of transparency that recognizes the freedom of individuals. For example, the use of data on same sex couples from the Census for advertising purposes is consistent with the liberal perspective because the use of such data recognizes the economic value and autonomy of the same sex couples, which can be as important a force as political or legal recognition. The liberal perspective gives equal weight to politics and markets in respecting the rights of individuals.

The autonomy perspective, on the other hand, recognizes that different realms provide different weight to aspects of individual autonomy. Being able to buy and sell is a different exercise of autonomy than being able to vote or legally marry. Therefore, the legal regime needs to recognize these differences and treat them accordingly. From the autonomy perspective, the Census example presented in the previous paragraph is a potential betrayal, allowing market recognition of same sex couples while denying political or legal recognition. Legal institutions should respond to this discrepancy accordingly by giving full weight to the autonomy of individuals.

Finally, the fairness perspective would consider the differing bargaining powers of actors in the political and market realm and scrutinize the commercialization of data from a distributive justice perspective. With respect to the Census example, the fairness perspective would treat the collection and use of the data as a potential exploitation of same sex couples and would support a ban on
the collection of the data. This conclusion contrasts with that of both the liberal
and the autonomy perspectives, which would both allow the collection and use of
the data. The liberal perspective would view the use of the data as a legitimation
of same sex couples and perhaps even a step towards formal political and legal
recognition. The autonomy perspective would be skeptical of it, but would
demand equal treatment of same sex couples in the economic, legal, and political
realms.

This section presents and assesses the legal structure of data
commercialization in light of this normative framework. As I show, the actual
legal structure governing data commercialization represents a compromise among
these three normative perspectives. In this way, data commercialization is made
possible while acknowledging competing normative perspectives. The analysis
presented in Section Three will provide the foundation for the licensing model for
data presented in Section Four, which is an attempt to reconcile the competing
normative forces by acknowledging them more completely.

A. Ownership and Data Commercialization

To what extent can data be owned? How does ownership in data relate to
ownership in information? If data and information can be owned, are their
restrictions on how each can be sold or shared? These are the questions addressed
in the rest of this subsection.

1. Intellectual Property Regimes

Four bodies of intellectual property law govern the ownership of data. I
discuss them in order of doctrinal and normative complexity; copyright, first, then
patent, trade secret, and common law misappropriation in that order. I end this
discussion with a short analysis of overarching First Amendment principles.

Copyright. The principal United States case directly addressing the
question of ownership of data is the 1991 Supreme Court decision in Feist v. Rural
Telephone Company. At issue in the Feist case was the copying of entries in a
telephone directory compiled and issued by the Rural Telephone Company in
Kansas. The Company issued separate telephone directories for each city in
Kansas; Feist took the separate directories for several of the cities and put together
a meta-directory that allowed individuals to search (in 1991 that meant manual and

56 See supra note 31.
visual searches rather than digital ones) the entire state. Ruling on the claim that Feist had infringed the Company’s copyrights in the directories, the Supreme Court emphasized two legal principles. First, one requirement for copyright protection is originality, which means that the purported author of the work was the original source for the protected material and that the author imbued the material with his or her own creativity. Second, with respect to compilations of data or facts, like a telephone directory, copyright protection does not extend to the facts or data compiled because they do not originate from the author. Instead, copyright protection extends to the creative selection, arrangement, and coordination of the facts and data by the author making the compilation. The Company was deemed not to have copyright in the alphabetical arrangement of names and other data in the phone directory, and Feist was free to copy that list of names and other data. The Court emphasized in its decision that copyright protection for database was thin, extending only to the particular selection, arrangement, and coordination of the facts and data, creatively chosen by the author.

The synopsis of the Feist decision conflates the terms facts and data. The Court in its opinion refers to facts and makes references on occasion to data. The Court’s rationale, however, applies equally to both facts and data, whatever the distinction. As the etymology of the word suggests, data is “given” by some source other than the person compiling data. Data exists presumably prior to the database and independent of the compiler. Therefore, data can be copyrighted because the first criterion for copyright protection is not met: data does not originate from the author. One still may ponder the relationship between facts and data. The data.gov website defines data as values assigned to certain concepts with the implication that these values are quantitative. Facts, however, can be qualitative as well as quantitative. The fact that it took Phileas Fogg seventy-nine days to circumnavigate the Earth in “Around the World in Eighty Days” contains both qualitative and quantitative facts. The quantitative fact would be data according to the working definition of data. This distinction may not be all the satisfactory because “values” can be qualitative as well. For this reason, I will leave others to make meaningful distinctions between facts and data. In this Article, I will use the term data synonymous with facts, and particularly in discussing copyright law.

57 See supra note 31 at 348.
58 See supra note 31 at 349.
59 See supra note 31 at 346 (referring to raw data as “wholly factual information not accompanied by any original written expression”).
The example of Phileas Fogg illustrates a more important than the semantic distinction between facts and data, the possibility that data could be fictional. Since originality for copyright purposes requires creativity, fictional data raises interesting possibilities for copyright protection. Copyright doctrine would treat the datum, “Phileas Fogg took seventy-nine days to circumnavigate the Earth,” in a number of ways.  

First, this datum is expression and as expression it can obtain some degree of copyright protection when fixed in a longer work, whether a paragraph like this one or a novel like Jules Verne’s. An additional requirement for copyright protection is that the expression be a recognized work of authorship, such as the compilation at issue in *Feist*. Other recognized works of authorship include literary works, musical works, and audiovisual works. A single sentence can constitute a literary work although the Copyright Office through its regulations has excluded short phrases, such as advertising jingles, from copyright protection in order to draw a distinction between copyright and trademark laws. A longer work, like this paragraph or a novel, containing the sentence would be a literary work, and the copying of the sentence would provide evidence of substantial similarity, the legal test for copyright infringement. In this way, copyright protection extends to the way in which data is expressed.

Second, copyright law extends only to the expression, and not to the idea being expressed. The data contained in a work of authorship would not be subject to copyright protection, as the Supreme Court made clear in *Feist*. Therefore, anyone can report that Phileas Fogg went around the world in seventy-nine days. Copyright law intrudes when the specific expression of that datum is copied. The ruling in *Feist* is just a specific example of this broader notion in copyright law. Under *Feist*, an individual is free to use the data but not the original selection, arrangement, and coordination of data, in other words the way in which the data is expressed through a database. The analogy would be to letters of the alphabet and a novel, or a note in a musical scale and a score. Copyright protection in a work does extend to building block components of the work, such as a letter, a note, or data.

---

62 See 17 USC 102(b).
63 See supra note 31 at 347.
Third, the challenging question is the treatment of created or constructed data. In the case of the Phileas Fogg example, the datum was initially created by Jules Verne in a book that is now in the public domain. To what extent would copyright protection extend to a fictional piece of data? At tension here are the principles of creativity, the lack of copyright protection for data, and the definition of work of authorship. Copyright is meant to promote creativity by giving a time limited protection to creative works for the purposes of management by the author and any eventual copyright owner. This principle pushes in favor of protection for created data. On the other hand, data is meant to be openly accessible and standing alone may not constitute a work of authorship. Add to this conundrum the observation that copyright’s directive is to promote progress in science, an old fashioned word for knowledge. By recognizing copyright protection for fictional data but not for non-fictional data, would not copyright law be promoting the creation of lies, falsehoods, and misconceptions? This last point is an intriguing one, but perhaps can be addressed by noting that value judgments and judgments of quality are not the domain of copyright law. Copyright law may protect falsehoods, but it is the bailiwick of other areas of law to police them. More troubling is the tension between creativity and access to data.

Case law subsequent to Feist has addressed the issue of constructed or fictional data, particular with respect to the creation of items like stock indices or sports statistic. In such cases, lower courts have made a distinction between data that would constitute a building block and data that would not constitute a building block. The former cannot be subject to copyright protection in order to protect accessibility, and the latter can in order to recognize creativity. The problem, of course, is determining when a piece of data is a building block. On this question, courts are not very helpful and tend to focus on the alterations made by the creator of the index or the sports statistic to the data. The analysis is analogous to the determination of what types of selection, arrangement, and coordination would constitute creativity. Another way to think about this analysis, within the language of this Article, is the distinction between data and information. Courts that

65 See, e.g., Belcher v. Tarbox, 486 F.2d 1087, 1088 (9th Cir. 1973) (“There is nothing in the Copyright Act to suggest that the courts are to pass upon the truth or falsity, the soundness or unsoundness, of the views embodied in a copyrighted work.”)
66 See Kregos, supra note 34 at 705.
67 Id.
68 See discussion, supra at text accompanying notes 12-13.
purport to protect non-building block data are effectively protecting information that is a creative interpretation of data. Stock indices and sports statistics are examples of how data is processed and analyzed to produce another work that represents an interpretation of the underlying data.ō Courts, therefore, are not protecting non-building block data but information, which represents, like a database, a way of processing the underlying data.69

The approach under Feist contrasts with that adopted by the European Database Directive in 1996.71 The Directive was enacted in response to concerns by members of the database industry that the thin copyright protection under Feist was not adequate to protect the economic interests of legal and financial database owners. The Directive extends protection for fifteen years to creators of databases which are the product of “substantial investment in obtaining, verifying, or presenting the contents of the database.”72 Database owners have the right to prevent imitation or alteration of the database and extraction of the data from the database. I discuss the Database Directive in more detail in Section Four of this Article. What is important at this stage is to identify the differences between the Directive and the approach pursuant to Feist. First, the Database Directive extends protection to the underlying data itself, to the extent the data is extracted from the database.73 Second, the Directive bases protection of databases on a “substantial investment” by the creator rather than any standard of creativity.74 Arguably, an alphabetic listing of names as in Feist would be protected if the compilation of such a list required substantial investment of time and money. Under Feist, a mere alphabetic compilation would never be protected by copyright law. The Directive therefore expands the scope of database protection and rights of the database creator. The United States had considered similar legislation in the 1990’s, but opposition from those seeking to protect access to data blocked the passage of analogous laws.75

Copyright law has several implications for the commercialization of data. Under the foundational principle of copyright, data would not be subject to copyright protection. Birthdays, nicknames, cities of birth, occupations, names of

---

69 See Kapes, supra note 32 at 1260-1261.
70 See Kregos, supra note 34 at 705; Feist, supra note 31 at 348-349.
72 Id. at art. 55.
73 Id.
74 Id.
friends, favorite colors, favorite musical groups, favorite books and the myriad other pieces of data that are stored on web sites or in nondigital form would constitute unoriginal items not subject to copyright, meaning they can be copied by anyone without fear of infringement. Expressions of any length, like posts or comments, would constitute copyrightable subject matter owned by the author of the expression. But these posts and comments can be gleaned and any data within them can be extracted without the risk of running afoul of copyright. Furthermore, many of these expressions may lose copyright protection because the expression and the idea expressed merge so that one cannot take the expression without also taking the data. For example, if a post says “I partied all night yesterday,” the data contained in the sentence cannot be separated from the sentence itself. Therefore, under the merger doctrine, the sentence itself loses copyright protection. For all these reasons, data is free of copyright protection and ready for anyone to commercialize.

The freedom of access to data under copyright law limits the use of copyright law as a means of extracting commercial value from data. If anyone can copy data, then no one would be willing to pay for data. The value comes from the construction of information and databases from the underlying data. What copyright law permits is an entrepreneur to process raw data to create an interpretative layer in the form of statistics or other composite information that synthesis or analyzes data into a valuable, and hence marketable, form. Such information could include summary statistics or reportable information that provides value for those unable to aggregate and synthesize the freely available data. In addition, value can be created through databases, which entail the compilation of data in creative ways. Databases do not involve the interpretation of data, necessarily, but do involve the packaging of data in ways that allow users to form their own interpretation and extraction of data. Commercialization of data very often involves the marketing of personal data in databases that serve to profile and target potential consumers of advertising and new products.

An interesting question arising from copyright law is what constitutes a database. Data extracted from, for example, a social networking site and arranged and coordinated creatively can constitute a database. But what about the web site itself? To what extent is all of Facebook a database? These questions have not

---

76 See, e.g, N.Y. Mercantile Exch., Inc. v. Intercontinental Exch., Inc., 497 F.3d 109, 116-17 (2007) (denying copyright protection to expressions of ideas that can only be expressed in a very limited number of ways, under the merger doctrine, such that the expression and the idea are so intertwined that they “merge” and are both ineligible for protection).

77 See Reichman & Samuelson, supra note 75 at 66-67.
been directly addressed, but they raise the possibility of a meta-database or a database that consists of other databases. The difficult possibility is that the layers of complexity made possible by meta-databases may undermine the underlying principle of access to data since at some point it may be difficult, if not impossible, to separate the copyright protected database from the freely accessible data. In this way, the owner of the meta-database may end up effectively obtaining ownership of the data and limiting access. Complicating the analysis is the number of ownership interests that might constitute the meta-database. Individual users may be responsible for owning portions of the database that they created while the creator of the web site itself will stake a claim in the aggregate of the databases. Conflicting and competing ownership interests may mitigate the ability of the meta-database owner to restrict access and thereby continue to keep data free from copyright protection.

**Patent.** The manipulation, extraction, and interpretation of data are the primary vehicles of extracting value through data commercialization. Copyright law protects the expressive uses of data through the creation of information and databases. But copyright law would not protect innovative ways of processing and using data. Such functional or utilitarian techniques are the subject of patent law. A survey of the United States Patent and Trademark Office database uncovers several hundred patents pertaining to the processing of data and the manipulation and use of databases. Patents, owned by companies like Google, Microsoft, Yahoo, and several smaller Internet based and computer outfits, cover search of databases, extraction of data, coordination and linking of databases, and other ways in which data can be arranged, coordinated, selected, and transformed. Professor Pamela Samuelson has argued that patents on data interfaces impede the interoperability of data and information and transfer mechanisms across systems, applications, and platforms. The following discussion explains how patent law is used to define and protect property rights in the data economy.

---

78 See id. at 114-115.
79 See id. at 103-104.
80 A search of patents in the USPTO online database uncovered 4343 patents which had the word “data” or “database” in the title and in the abstract. The search was conducted on February 10, 2011.
81 See, e.g., “Method and Apparatus for Exchanging Data with a Database,” Patent # 7877417 (January 25, 2011)(patent assigned to Microsoft covering invention that allows for extraction of data from database of any structure); “Searching Structured Geographic Data,” Patent # 7836085 (November 16, 2010)(patent assigned to Google for searching and extracting data from multiple datasets); “Expanding a query to include terms associated through visual content,” Patent # 7882124 (February 1, 2011)(patent assigned to Yahoo! for searching databases containing visual content).
A patent protects a novel, useful, and nonobvious invention. An invention is defined in the United States Patent Act as a process, machine, manufacture, or composition of matter.\textsuperscript{83} Case law establishes three exclusions from the definition of invention: abstract ideas, laws of nature, and natural phenomena.\textsuperscript{84} Data itself would not be patentable since it does not fall into any of these four categories of invention. Processed data in the form of information or a database would also not fall under patentable subject matter.\textsuperscript{85} However, tools to extract, aggregate, manipulate, or process data would be considered as a patentable invention. A general purpose machine, such as a computer, can be patented.\textsuperscript{86} Similarly, a special purpose machine, such as a heart monitor, that collected, extracted or analyzed special types of data, such as health indicators, can be patented.\textsuperscript{87} Furthermore, processes for manipulating or transforming data that are not necessarily tied to or operationalized through a machine can be patented. So methods for search or for data extrapolation can be patented.\textsuperscript{88} A patent is obtained after an application and review process with the United States Patent and Trademark Office.\textsuperscript{89} The administrative process checks to see if the invention is novel, nonobvious, useful, and a patentable invention.\textsuperscript{90} Once granted, the patent allows the owner to exclude others from using, making, selling, or importing the patented invention for a period of twenty years from the date of application.\textsuperscript{91}

A critical question with any type of patent is that of validity. The USPTO grant of a patent creates a presumption of validity, but this presumption can be rebutted in the context of patent litigation.\textsuperscript{92} A challenger to a patent would have to show that the invention is not subject matter covered by the statute or that the invention is not novel, nonobvious, useful, or disclosed in the patent specifications. Novelty and nonobviousness entails an analysis of the prior art leading up to the invention while usefulness and disclosure requires a consideration of the patent application and the nature of the invention. As for patents pertaining to data, such as the ones discussed above, the critical concerns for patent validity would be whether a particular invention is novel and nonobviousness and whether a

\begin{footnotesize}
\begin{itemize}
\item \textsuperscript{83} 35 USC 101.
\item \textsuperscript{84} See Diamond v. Chakrabarty, 447 U.S. 303, 309 (1980).
\item \textsuperscript{85} See Baker v Selden, 101 U.S. 99 (1879).
\item \textsuperscript{86} See In re Alappat, 33 F.3d 1526, 1545 (Fed. Cir. 2003)
\item \textsuperscript{87} Id.
\item \textsuperscript{88} See In re Alappat, 33 f.3d 1526 (Fed. Cir. 1994); State Street Bank & Trust Co. v. Signature Financial Group, Inc., 149 F.3d 1368 (Fed. Cir. 1998).
\item \textsuperscript{89} 35 USC 131-13.
\item \textsuperscript{90} 35 USC 101-103.
\item \textsuperscript{91} 35 USC 271.
\item \textsuperscript{92} 35 USC 282.
\end{itemize}
\end{footnotesize}
particular invention would be a process, machine, manufacture, or composition of matter.  

As for novelty or nonobviousness, the validity question would rest on the prior art in the relevant field within which the data related invention falls. The novelty inquiry rests on a fairly broad understanding of the prior art; it can be from any field. The nonobviousness inquiry rests on a narrower scope of the prior art; it can come from an analogous field to the invention at issue. For the most part, the prior art for data related inventions will be from the fields of computer science or electrical engineering. Each of these fields will present the state of knowledge on electrical processes, computer flow charting, and data processing that will aid in assessing the invention. The novelty analysis will focus on whether an exactly identical invention already existed in the prior art. The nonobviousness analysis will focus on whether the differences between the patented invention and the what was known in the prior art would be obvious to a person having ordinary skill in the art. In other words, do the differences constitute trivial variations or substantive progress in the field? On this last inquiry, data related patents may be subject to vulnerability as there may be very few new methods of processing data that are truly innovative. To emphasize once again, the specific inquiry will depend on the field and the scope of the prior art.

The Supreme Court’s recent decision in *Bilski v. Kappos* has direct relevance to data related patents. At issue in *Bilski* was what types of processes would constitute a patentable invention. The USPTO rejected a patent application disclosing a method for hedging financial risk. The disclosure did not limit the method to a particular machine or physical context. The patent examiner, following the approach of the European Patent Office, rejected the application on the grounds that there was no technical effect disclosed in the invention, meaning that the disclosure merely described a series of steps, but did not disclose any operational or engineering steps. The Board of Patent Appeals and Interferences, the first level of review within the USPTO, affirmed the rejection.

---

93 See e.g. In re Bilski, 545 F.3d 943, 976-977 (2008)(Judge Newman’s dissent identifying issues of novelty and nonobviousness in addition to patentable subject matter in addressing patenting of processes).
94 35 USC 102-103.
95 See, e.g., In re Lowry, 32 F.3d 1579 (Fed. Cir. 1994)(patent for invention involving data processing system).
99 Id. at 3221.
100 Id..
but based the rejection on the grounds that the invention was an abstract idea, as opposed to a concrete application.\textsuperscript{101}

The United States Court of Appeals for the Federal Circuit reviewed the rejection and upheld it but on substantially different grounds. Reviewing the Supreme Court precedent on the patentability of processes, particular software related processes, the Federal Circuit held that not all processes are patentable.\textsuperscript{102} The court held that in order for a process to be patentable it either had to be machine based or constitute a transformation.\textsuperscript{103} The difficult part of this “machine or transformation” test is discerning when a process constitutes a transformation.\textsuperscript{104} Partly, a transformation involves something physical, such as a change from one physical state to another as in a chemical or biological process.\textsuperscript{105} But a transformation could also involve a change that is nonphysical, such as the transformation of data from one form to another through mathematical or other operations. In case of the hedging patent at issue in Bilski, the Federal Circuit ruled that the invention did not constitute a patentable process because the hedging method at issue did not involve a machine and did not involve a transformation of data from one state to another.\textsuperscript{106} The process disclosed was purely mental or cognitive.\textsuperscript{107}

The Supreme Court granted certiori on Bilski's appeal from the Federal Circuit and affirmed the rejection, but not the court’s reasoning. The Supreme Court held that processes in general were patentable unless they constituted an abstract idea, a law of nature, or natural phenomena.\textsuperscript{108} The machine or transformation test was not the sole test to determine a process would qualify as an invention, but was one of many possible tests.\textsuperscript{109} Finally, the Supreme Court affirmed the rejection on the ground that the hedging method was an abstract idea and therefore excluded from the meaning of an invention.\textsuperscript{110}

In light of Bilski, data related patents may face some vulnerabilities. The Bilski opinion is an opaque one, revealing certain fault lines in the Roberts court.

\textsuperscript{101} Id.
\textsuperscript{102} See supra note 93 at 952-953.
\textsuperscript{103} Id. at 956.
\textsuperscript{104} Id. at 957.
\textsuperscript{105} Id .
\textsuperscript{106} Id. at 958.
\textsuperscript{107} Id. See, also, In re Comiskey, 554 F.3d 967 (Fed. Cir. 2009)(rejection of arbitration method patent as invalid subject matter because the method was an abstract idea and mental step).
\textsuperscript{108} Bilski, supra note 98 at 3222.
\textsuperscript{109} Id. at 3228.
\textsuperscript{110} Id. at 3226.
The unanimous decision speaks towards a broad reading of “process”, implying that most data related patents would be found valid.\textsuperscript{111} However, four of the nine justices (Stevens, Sotomayor, Breyer, and Ginsburg) would have found an exception for business method patents from patentability.\textsuperscript{112} Since many data related patents are types of business method patents, a broad exclusion for business methods would make data related patents vulnerable.\textsuperscript{113} It is unlikely, however, that the Supreme Court would recognize such a broad exclusion, especially now that Justice Stevens has retired.\textsuperscript{114} A more realistic threat to data related patents would be the interpretation of Bilski by the USPTO and the Federal Circuit. The Supreme Court recognizes limitations on the meaning of invention and did endorse the machine or transformation test as one test for patent validity. Subsequent to the Bilski decision, the USPTO issued proposed rules for analyzing patentable subject matter.\textsuperscript{115} Future developments will show how far these limitations will be taken. What is the case, however, is that patents serve as an important type of intellectual property for the protection of methods of data processing which complement the protection to information and databases protected by copyright.

**Trade Secret.** Inventions related to data and databases, information, and databases, and even data itself can be protected as a trade secret. Complementing both patent and copyright, trade secret law provides a comprehensive form of intellectual property, grounded mostly at the state level and through criminal enforcement by the Department of Justice at the federal level.\textsuperscript{116} Given its breadth, trade secret may serve as a substitute for patent and copyright. Despite the breadth of trade secret law, intellectual property owners may prefer patent and copyright to trade secret protection for many reasons including access to federal courts and the absence of strong defenses to infringement like reverse engineering, which is a critical feature of trade secret law.\textsuperscript{117} Stated succinctly, trade secret offers broader coverage of potentially perpetual duration but subject to stronger protections for alleged infringers, as compared to patent and copyright.

Trade secret protects any type of “information” whose value comes from the fact that it is not generally known and easily discernible and that a competitor does

\textsuperscript{111} Id.
\textsuperscript{112} Id. at 3229.
\textsuperscript{113} Id.
\textsuperscript{114} Id. at 3227. (Court hesitant to adopt categorical rules).
\textsuperscript{117} See Kewanee Oil Co. v. Bicron Corp., 416 U.S. 4704, 481 (1974)(comparing trade secret with other intellectual property law).
not have access to it.\textsuperscript{118} The word information is used in most trade secret statutes and should not be confused with the usage in this Article. The concept of information is broad-ranging and includes processes, formulas, and data.\textsuperscript{119} For example, customer lists are protected as a trade secret in all jurisdictions and individual pieces of data, such as a price or the name of a customer, can be protected as a trade secret. Therefore, unlike with patent and copyright, trade secret can protect data itself and not just simply functional methods for processing data or expressions interpreting or compiling data.

Although trade secret covers wide subject matter, including data, there are several limitations on the role of trade secret in commercializing data. First of all, what counts as a trade secret depends on value in a commercial or market context. Although anyone who breaches a trade secret, whether a direct competitor or not, can be liable for trade secret misappropriation, the protection of trade secret law extends only to commercially valuable secrets. Issues of privacy and security, which are also sources of value, are not relevant to trade secret protection.\textsuperscript{120} Therefore, trade secret law does not protect privacy interests in data but only commercial interests.

Second, trade secret protection does not extend to readily ascertainable information.\textsuperscript{121} Data like birthdates, names, addresses, and market variables may be excluded from trade secret protection if they are each ascertainable from independent sources. Owners and operators of social networking sites would not be able to protect such readily ascertainable information as a trade secret. The exclusion would extend to data contained in items posted by users. If such posts are public and readily searchable, then their data content would be excluded from trade secret protection.

Third, trade secret protection requires affirmative steps by the purported trade secret owner who must take reasonable steps to maintain the secrecy of the

\textsuperscript{118} The Uniform Trade Secret Act (UTSA) defines a trade secret as:
\[\text{I}nformation, \text{including a formula, pattern, compilation, program, device, method, technique, or process, that:} \]
\[\text{(i) derives independent economic value, actual or potential, from not being generally known to, and not being readily ascertainable by proper means by, other persons who can obtain economic value from its disclosure or use, and (ii) is the subject of efforts that are reasonable under the circumstances to maintain its secrecy.}\]

\textsuperscript{119} See, e.g., Ruckleshaus v. Monsanto, Co., 467 U.S. 986, 1003 (1984)(finding that environmental data submitted to government could be protected as a trade secret).

\textsuperscript{120} See, e.g., Dow Chemical Co. v. United States, 476 US 227, 239 (1985)(trade secret not relevant to reasonable expectation of privacy analysis);

\textsuperscript{121} See UTSA, supra note 118.
protected information. Such reasonable steps could include technological measures (for example, encryption or password protection) or contractual measures such as nondisclosure agreements or covenant not to compete. Reasonable steps could also include the creation of special relationships which would impose a duty of confidentiality on the parties. This requirement makes the protection of trade secret particularly difficult for the use of trade secret law in the commercialization of data. Often data will be created by many users independent of any relationship with the trade secret owner. Therefore, it will be impossible for the trade secret owner to maintain the secrecy of the generated data. Consider social network sites where anyone can join and posts can be readily searched and copied. A term of use for the site that imposed confidentiality requirements on the data and information generated on the site may be difficult to enforce (and would face the same technical and social challenges as protection for personal privacy). While reasonable steps are a loose requirement and may in some instances be a formal one, the vulnerability of data generated by multiple users would be grounds for questioning whether this requirement for trade secret protection has been met.

In theory, trade secret law can protect data. In practice, however, such protection can be difficult to obtain and sustain. As a consequence, trade secret protection may be limited to protection for information and databases as well as for methods of processing and manipulating data, the same subject matter as copyright and patent. In this way, trade secret law is an important complement to federal intellectual property protection and may aid in data commercialization through longer duration of protection.

**Misappropriation.** Despite its uncertain roots, the common law tort of misappropriation provides a quasi-property right for certain types of data and information. Traced to the Supreme Court’s decision in *INS v. Associated Press*, the tort of misappropriation allows the aggregator, collector, or disseminator of certain types of data or information to prevent a competitor from improperly accessing and using the data or information for a limited period of time. The impropriety is based on the actions of the party who has allegedly performed the misappropriation and focuses on the extent to which that party has attempted to

---

122 See UTSA, supra note 118.
124 248 U.S. 214 (1918).
125 See National Basketball Association v. Motorola, 105 F.3d 841 (2nd Cir. 1997)(presented a multifactor test as basis for valid misappropriation claim, where factors included timeliness of information, effort expended to obtaining information, free riding in copying information, and competition between plaintiff and defendant).
free ride off the efforts of claimant.\textsuperscript{126} Courts typically focus on whether a party has simply copied someone else’s data or information or has added value in how the data or information is generated or used. In this way, the common law tort of misappropriation extends trade secret law to protect items that would not qualify as a trade secret.\textsuperscript{127}

Timing is key to the tort of misappropriation. The data or information protection must fall into the category of “hot news,” in other words data or information that is time sensitive and whose commercial value expires quickly because it is readily leaked to the public or readily discoverable by independent sources. In the \textit{Associated Press} case, protection extended to news itself that was gathered from the battlefields of Europe.\textsuperscript{128} In other cases, hot news has included sports scores and market prices.\textsuperscript{129} Unlike trade secret law, the tort of misappropriation is designed to protect the labor of the creator in collecting, aggregating, and manipulating data.

What distinguishes the tort of misappropriation from the other types of intellectual property described in this section is its scope. The tort protects specific time sensitive data that may not qualify as a trade secret, a patentable invention, or copyrightable expression. Furthermore, the theory of the tort is the protection of labor invested into the generation or manipulation of the data, much as with the European Database Directive.\textsuperscript{130} Finally, the tort creates a quasi-property right since it only provides protection against misappropriation by a direct competitor. As applied to the commercialization of data, the tort provides highly specialized protection that might fill a niche left open by copyright, patent, and trade secret.

\textbf{The Normative Foundations of Data Commercialization.} The ownership structure of data reveals an overlap of the three normative perspectives described above. The emphasis on property protection to promote creation as well as management is consistent with the liberal perspective. Allowing some property protection for information, processes, and secrets promotes participation in both the market and political spheres. Furthermore, the construction of data and information as property makes no distinction between the political and economic

\textsuperscript{126} Id. at 845 (discussion requirement of free riding).
\textsuperscript{128} See supra note 124 at 231.
\textsuperscript{129} See Motorola, supra note 125 at 854. For a successful misappropriation claim involving stock indices, see Board of Trade v. Dow Jones & Co., 456 N.E.2d 84 (1983)
\textsuperscript{130} See supra note 71 and discussion infra at Section IV.A.3.
uses of information. In addition, the relevant open access given to data under the various regimes reflects the existence of a public sphere that is exempt from both economic commercialization and private secrecy. Access to data reflects a form of transparency fundamental to both politics and markets. In fact, the only area where secrecy of data is countenanced is when such secrecy permits individual economic gain, a result consistent with the liberal perspective on economics and politics.

The autonomy perspective would interpret the ownership structure in a different light. The relatively free accessibility of data reflects a domain where individuals are free to operate without the intrusion of commerce or politics. The commercial sphere encloses a domain of data, which can be understood as refined and processed data, which permits commercial exploitation. Concerns of privacy and noncommercial space seem largely absent from the ownership structure, but may be implicit both in the scope of the public domain and in the distinctions among the different ways in which data are protected. Copyright and patent, and to a certain extent trade secret and misappropriation, protect highly refined data and the tools for refining data. The result is a middle ground between truly raw data and highly refined data where personal space operates through protection of individual autonomy. This middle ground will be explored more in the section below on sharing and selling data. One potential conflict between the liberal and autonomy perspectives is that this middle ground simply may not exist for the liberal perspective as all data is divided in that which is raw and open and that which is refined and proprietary.

Finally, the fairness perspective would cast this ownership structure in terms of competing sources of power, both political and economic. Allowing data to be openly accessible creates a zone free of the exercise of commercial and political power. Data transparency serves to ensure equal access and a commons free of proprietary influence. A proprietary zone is appropriate to allow individuals to process and hence commercialize data, but this zone should be narrowly construed. The fairness perspective would deviate from the liberal and autonomy perspectives on the need for law to protect the distribution of data and information among market and political agents. On this point, the fairness perspective would ask more of the ownership structure than I have described at this point. It would call for more internal limiting doctrines (such as fair use and the first sale doctrine) as well as external limiting doctrines (such as laws protecting privacy, unfair

131 See Pamela Samuelson, The U.S. Digital Agenda at WIPO, 37 Va. J. Int'l L. 369, 381 n.74 (1997) (discussing fair use and first sale doctrines as important limits on ownership rights over information serving to protect access).
competition, and misuse of personal information). In short, the fairness perspective would ground greater data transparency as well as limits on data commercialization.

An important constitutional background to the commercialization of data is the protection given to freedom of speech under the First Amendment of the US Constitution and comparable protections in other jurisdictions. The First Amendment is in the background interjecting in only special cases. From the liberal perspective, the ownership structure complements the First Amendment by creating a market for expressions that complements the marketplace of ideas. Economic freedom and political freedom go hand in hand to protect full citizen participation in the economic and political realms. The autonomy perspective would see the First Amendment limited by concerns for privacy and security in the middle ground that arises between raw data and highly refined data. In the two tails, however, the First Amendment would serve as a complement to the ownership structure, much as with the liberal perspective. Finally, the fairness perspective would use the First Amendment as a limiting principle to protect against power imbalances in the various spheres in which data exists. Speech serves to protect politically or economically disadvantaged groups from the misinterpretation and misuse of data and the generation of information that can be misleading and therefore politically and economically harmful.

In short, the ownership structure over data is both descriptively and normatively complex. Each of the normative perspectives plays a role in the governance of data. A similar complexity can be seen in the government ownership of data and in the notion of open records.

2. Government Data and Open Records Acts

---

133 See, e.g., IMS Health Inc. v. Sorrell, 630 F.3d 263 (2nd Cir. 2010), cert granted by Sorrel v. IMS Health Inc., 131 S. Ct. 857 (2011)(First Amendment challenge to state statute prohibiting the sale, license or exchange of pharmaceutical data).
134 As examples of this view of the First Amendment, see Eldred v. Ashcroft, 537 U.S. 183 (2003)(Court finding narrow First Amendment limits on copyright law); see Sorrel, supra note 133 at 274 (affirming First Amendment claim to allow for commercialization of data).
135 See, e.g., IMS Health Inc. v. Ayotte, 550 F.3d 42 (1st Cir. 2008)(upholding statute limiting commercialization of pharmaceutical data against First Amendment challenge).
137 See, e.g., Houchins, supra note 136 at 12 (discussing interests in protecting inmate population from news reporting and gathering).
Governmental bodies, including agencies, generate data and information. For the most part, the rules of intellectual property law apply to such data and information although the particular context of government shapes the particular application of the intellectual property rules. For example, federal government bodies cannot own copyright in legal materials under the United States Copyright Act unless the materials are created by an independent contractor who transfers copyright to the government. However, the treatment of databases created by the federal government would be different from legal materials, such as statutes, regulations, and case opinions. Such databases would be subject to copyright protection. There are no statutory restrictions on ownership of copyright by state governments although some jurisdictions have created restrictions for state legislative and regulatory materials. As far as patent law, governments have not obtained patents in data or information related inventions although they could. However, governments do own trade secrets in data and information, which has been the source of controversy for proponents of open government. Finally, misappropriation would exist as a cause of action for governmental entities although no reported cases illustrate such claims.

Government owned data and databases pose challenges for public accountability and access. All fifty states have enacted an open records act, which make governmental entities the custodian of records with the obligation of disclosing certain records upon a proper request. Open records acts are often considered the product of Progressive Era legal reform, designed to make government more accountable and accessible to the public. However, many states adopted Open Records Acts prior to the period known as the Progressive Era, roughly the period from 1890 to 1920. The State of Wisconsin, for

---

138 For an analysis of exclusions from intellectual property protection of legal materials, see Shubha Ghosh, Copyright as Privatization: The Case of Model Codes, 78 Tul. L. Rev. 653, 663-668 (2004) (analyzing importance of excluding statutes, regulations, and case law from copyright protection from a democratic governance perspective).


142 Stewart, supra note 141 at 266.
example, often associated with quintessential Progressive Era reforms, enacted its Open Records Act in 1850, shortly after its inclusion in the United States.\textsuperscript{143} The first version of the act was an absolute open statute, meaning that most of the restrictions on access were procedural ones.\textsuperscript{144} Over time, however, courts placed reasonable limits on access to protect interests such as privacy, property, or security.\textsuperscript{145} The current version of the Open Records Act has been the subject of controversy and will serve as a case study in Section Four.

An analysis of the current version of the Wisconsin Open Records Act highlights many of the themes of this Article and raises additional questions. The statement of purpose enunciated in the Act illustrates both its bold spirit and some potential tensions:

it is declared to be the public policy of this state that all persons are entitled to the greatest possible information regarding the affairs of government and the official acts of those officers and employers who represent them. Further, providing persons with such information is declared to be an essential function of a representative government….\textsuperscript{146}

The provision concludes by creating a “presumption of complete public access” with access being denied “only in an exceptional case.”\textsuperscript{147}

Terminology is critical to the application of the Open Records Act. The purpose statement pertains to access to “information,” a term that is not defined in the statute. The word data is not used in the statement of purpose although it does arise in another section discussed below.\textsuperscript{148} Presumably, the word information has its ordinary meaning and not the specialized meaning examined in this Article. This ordinary meaning would include not just interpretations of raw data, but facts as well as they arise in governmental proceedings. The key concept, however, is neither information nor data, but a record.\textsuperscript{149} Needless to say, the purpose of the

\textsuperscript{143} Kidwell, supra note 141 at 1028.
\textsuperscript{144} Kidwell, supra note 141 at 1030.
\textsuperscript{146} Wisc. Code 19.31 (1999).
\textsuperscript{147} Id.
\textsuperscript{148} See definition of record, infra.
\textsuperscript{149} See Kidwell, supra note 141 at 1026-1027 (drawing distinction between record and copyrighted material).
Open Records Act is not simply access to information, but recorded information. The definition of record is pivotal to the application of the Act.

There is no obligation to create a record under the Open Records Act. But if a record is kept, then it is subject to accessibility. A record is defined as

any material on which written, drawn, printed, spoken, visual or electromagnetic information is recorded or preserved, regardless of physical form or characteristics, which has been created or is being kept by an authority.¹⁵⁰

The authority referred to in the definition is a governmental authority. As we will see in the case study in Section Four, this authority can include a non-governmental authority to whom power has been delegated by a governmental authority. The definition speaks broadly in terms of the medium that would constitute a record and the potential types of information that would constitute a record. However, the definition excludes copyrighted and patented materials and published materials that are on sale or available in a public library from the definition of a record.¹⁵¹ Also excluded are drafts, notes, and other preliminary materials used to create a record.¹⁵² These exclusions protect certain property and economic interests as well as the ability of an authority to create a record without having to disclose the thought process leading up to the record’s creation. A record is a statutorily recognized and official version of the information (and presumably data) that the Open Records Act makes accessible to a requesting public.

Exclusions also exist for requests for records under the Act. Specifically, the statute excludes trade secrets, computer data, and proprietary data from a record request.¹⁵³ This exclusion applies to the request, but not to the definition of record, which means that the authority subject to a request can deny access if the request is for what is deemed to be a trade secret or restricted data. This distinction is important because the inclusion of these items as a record means that the authority must serve as a custodian and maintain these items as a request. There is, however, authority to exclude these items from a record request.

¹⁵¹ Id.
¹⁵² Id.
Open records acts provide an important comparison with data.gov. Although both ostensibly serve the goals of public access, each illustrates different approaches. Open records acts serve as a means to create an official recording of governmental acts that are subject to the governmental stewardship and to citizen requests for access. The acts serve an archival purpose that documents the actions of governmental entities. A website like data.gov also serves an archival purpose of sorts, but its design is to allow individuals to access and use data that is collected or held by the government. Neither requires or mandates the creation of a record or the accumulation of data. The presumption is that these activities will occur anyway. The two mechanisms are created in order to allow access of sorts to the records and data created. The open records act can be compared to a museum or a library whose collection can be accessed under certain circumstances. A website like data.gov, by comparison, is like the Exploratorium, which allows attendees to actually touch, play with, and otherwise use the items collected. Section Four will explore these differences in greater detail in the context of case studies.

Government ownership of data has three different dimensions, each reflecting the normative complexity of the legal structure of data ownership: (1) the role of government as market actor; (2) archiving data and information; and (3) transparency and openness. Each of these three dimensions reflects the normative complexity in its own way.

Government control of data reflects the government’s role as a market actor, both through intervention in market activity and through participation. As an intervenor, the government obtains data from regulated entities but also manipulates and interprets such data through the regulatory process. When the government is an active market participant, through state owned enterprises, the government also participates in the generation and accumulation of data. Such a role is consistent with the liberal perspective that would consider strong overlap between the market and political realms. The autonomy perspective would see conflict, however, especially if the government’s role as regulator feeds into its role as market participant, perhaps through the use of data obtained through the regulatory process by state owned enterprises. Finally, the fairness perspective would be concerned with the undue influence of the state in the private realm through its coercive and potentially monopoly power. As the case studies in Section Four will show, these distinct normative perspectives can overlap and conflict in practice.
The archiving dimensions of open records acts also reflect this normative tension. The liberal perspective would consider the archiving function part of the dual market and political realms of liberal society. Open records provide a source of information that can aid in political deliberation and market participation. Arguably, and this is a point that will be examined more in Section Four in a case study, the web site data.gov exemplifies the liberal perspective as data accessibility provides a way to both gain knowledge useful in the political realm with minimal limits on commercialization. But both the autonomy perspective and fairness perspective would support a cautionary take on data archiving. Individual autonomy involves remembering as well as forgetting the past, and the archiving of data provides a trail of one’s past that one might prefer to expunge. Similarly, the fairness perspective would draw our attention to power imbalances in how data is recorded and stored. The rich can have records expunged or controlled while the poor may not have this capability.

Finally, the normative tensions are apparent in the transparency and openness provided by both open records acts and web sites like data.gov. The liberal perspective would herald transparency, echoing the sentiments of Justice Brandeis on sunlight as the best disinfectant. Once again, the autonomy and fairness perspectives would invite caution. Transparent to whom and for what purposes would be the questions raised under these two perspectives. The autonomy perspective would be concerned with who is allowed access to records and for what purposes. The fairness perspective would be concerned with equality issues that are overlooked in open records acts, which emphasize values of freedom over equal distribution of resources for access to and use of records that may contain sensitive data and information.

Government data ownership is a critical part of the legal ownership of data. This section has highlighted some of the critical issues that will be explored in greater depth through the case studies in Section Four, particularly the case studies on access to tax records and real estate assessments and on data.gov.

B. Selling, Sharing, and Data Commercialization

Legal ownership of data provides a necessary condition for transacting in data either through sharing or market transactions, but such transactions may be further limited by legal rules. Such legal rules are designed to protect interests of transacting parties. The legal structure of transactions maps onto the rules of
contract, intellectual property licensing, tort law, and other regulatory fields. But there are several specific structures worth highlighting here.

First, United States law allows for restrictions on the use of data through contract law. In *ProCD v. Zeidenberg*, the Seventh Circuit held that a seller of a database can restrict access to data through a contractual limitation.\(^{154}\) Such a contract term was found not to be preempted by federal intellectual property law and was held to complement the interests of intellectual property holders. This decision has been a controversial one since it was first handed down in 1996, but with some narrow exceptions regarding contractual formation, the ruling is generally accepted.\(^{155}\) The rationale behind this contractual protection for data is that the restriction on data access applies only to the parties to the specific transaction and therefore does not conflict with the copyright principle that data cannot be copyrighted. Furthermore, the contractual provision is deemed to protect the expectations of the party allowing the seller to turn the data into a marketable commodity from which economic value can be extracted beyond the property right in the database. Whether such a rationale is consistent with intellectual property policy has been a central controversy. Even if the restriction is between the transacting parties, mass market licensing affectively binds any purchaser to the term restricting data. Therefore, what seems ostensibly to be a contract right becomes a property right, inconsistent with copyright law. Admittedly the remedies may be different for a contract breach than for copyright infringement, but purchasers may not in practice feel the difference between being bound by contract and being bound by copyright law. In addition, the use of arbitration to enforce the term may limit the difference in remedies and thereby further diminish the difference. An open and critical question, which will be taken up in Section Four, is the extent to which technological protections, such as encryption, can be used to protect data.

However, legal rules also attempt to protect against privacy invasion and interference with personal interests. As many scholars have noted, such protections are piecemeal, often dependent on the type of data with the Health Insurance Portability and Accountablity Act of 1996 (HIPAA) regulations providing one example of detailed government regulations protecting privacy interests in health data.\(^{156}\) The tort of intrusion generally protects against invasions

\(^{154}\) 86 F.3d 1447 (7th Cir. 1996).


\(^{156}\) See Solove, supra note 15 at 69-70 (analyzing HIPAA).
of one’s reasonable expectation of privacy, and the related tort of false light prevents reputation harming uses of data. The right of publicity, one of the original common law torts of privacy, has taken on a life of its own and protects against the unauthorized commercial exploitation of one’s public persona. For all of these torts, the First Amendment places some limitations on tort claims against reporting or communication of private data and information. Finally, the Fourth Amendment would place limits on the government, whether acting in its crime enforcement or its regulatory capacity, from unwarranted searches and seizures and data and information obtained directly or indirectly from such searches or seizures.

With advances in genetic technologies, claims against the misuse of personal genetic information have grabbed the attention of scholars and policymakers. Statutes against genetic discrimination prevent unequal treatment in the use of genetic data.\textsuperscript{157} Collection of tissue samples and cell cultures are not governed by property law,\textsuperscript{158} but would be regulated by tort law. The famous case of John Moore whose tissue line was extracted by medical researchers at UCLA Medical School brought harvesting of what is in effect personal genetic information under legal scrutiny.\textsuperscript{159} Although one cannot claim a property right per se in one’s body or genetic materials, one can bring claims of lack of informed consent and breach of fiduciary duty against researchers who misuse such materials. Such potential claims shape transactions in genetic data and the increased use of informed consent in hospital and medical research settings.\textsuperscript{160} The case of Henrietta Lacks recently received scholarly and media attention.\textsuperscript{161} A cell line obtained during her treatment for ovarian cancer had been used by several generations of medical researchers and students after her death. Her example illustrates broader trends in personalized medicine and the critical role of personalized medical and genetic information in the context of the patenting and marketing of medical diagnosis and treatment methods.\textsuperscript{162}

The legal structure of data based transactions illustrates the same normative complexity as the legal structure of data ownership. The three normative perspectives surface and justify different parts of the legal structure. The liberal perspective supports a contracting regime to protect data through contractual


\textsuperscript{158} See Washington University v. Catalona, 490 F.3d 667, 674 (8th Cir. 2007)(biological materials treated as inter vivos gifts from research participant to university).

\textsuperscript{159} Moore v. Regents of the University of California, 793 P.2d 479 (1991).

\textsuperscript{160} See Skloot, supra note 8 at 207-211 (describing benefit sharing agreement between hospital and patient permitting harvesting of patient’s biological materials).

\textsuperscript{161} See Skloot, supra note 8 at 305-306.

\textsuperscript{162} See Ghosh, supra note 8 at 107.
restrictions, as we will see in greater depth in the case study on open records and real estate assessment data in Section Four. The autonomy and fairness perspectives provide support for several limitations on data based transactions, especially through protections for privacy and for the exploitation of genetic data and information. Section Four explores these tensions through the analysis of three case studies, each bringing together the legal structure of ownership and transactions presented in this section. These case studies will be instrumental in building the open source licensing model for data transactions presented at the end of Section Four and which brings together many of the themes and arguments of this Article.

IV. Towards a Licensing Model for Data

This section brings together the economic, legal, and normative issues raised by the commercialization of data through three case studies. Each case study demonstrates how goals of transparency and openness come into tension with the pursuit of commercialization. The analysis of the three case studies provides a foundation for disseminating and commercializing data while allowing restrictions on use that preserve values of autonomy and fairness through what can described as open source licensing terms.

A. Three Case Studies: Real Estate Data, Government Data, and European Databases

This subsection presents three case studies. The first has to do with access to property assessment data collected by municipalities through private corporations. The case study focuses on transparency of data collected by the government and the use of such data for commercial purposes. The tension between governmental transparency and commercialization illustrates how state and markets, public and private institutions, overlap in the creation and management of data. The second case study examines the interaction between public and private under access to federal government data and transparency at the federal level. This case study shows how regulation controls access and use of federal data. Finally, the third case study examines the treatment of data and databases under the European database initiative and shows how a regime designed for the proprietary treatment of data and databases has shifted to one which provides limits on private ownership. Together the three case studies illustrate the economic, legal, and normative issues framed in the previous sections.
Furthermore, these case studies provide a foundation for the licensing model to be presented in the final subsection.

1. Real Estate Data

In 2002, WireData, a private company providing information for realtors on real property that is up for sale, submitted a request under Wisconsin’s Open Records Act on various municipalities in the State of Wisconsin for tax and real property assessment data. The goal was to include these data in online real estate listings so that potential buyers would have more information for properties that might be of interest for purchase. Some municipalities informed WireData that they did not have the data which was collected by Assessment Technologies, a private company that compiled and archived the assessment data. WireData requested the data from Assessment Technologies, which denied the request on a number of grounds including the rationale that the Open Records Act did not apply to private entities and that even if the Open Records Act applied, the assessment records were exempt from the Act because they were protected by copyright law. WireData sought a declaratory judgment on the copyright status of the real estate data that it was seeking. A magistrate judge on the Seventh Circuit ruled in favor of Assessment Technologies. WireData appealed successfully to the Seventh Circuit, which ruled in an unanimous opinion by Judge Richard Posner that copyright law protected original databases but not the data of which the database is composed. The Seventh Circuit held that the use of copyright law to deny access to unprotected data constituted copyright misuse, an equitable doctrine that limits the right of a copyright owner that attempts to extend the scope of copyright. The effect of the appellate ruling was that Assessment Technologies could not refuse WireData’s request on the grounds that the records were protected by copyright.

The Seventh Circuit ruling set the stage for the final acts in the drama of WireData’s search to obtain what it viewed as the holy grail: real estate data that could be electronically manipulated. WireData had requested electronic copies of the assessment data so that they could be readily entered into its web site of real estate information. In light of the Seventh Circuit ruling, Assessment Technologies decided to comply with the request for electronic copies of the data by providing electronic files of the assessments in a .pdf format. While this format was

164 Id. at 743.
165 Id. at 745.
166 Assessment Technologies of WI, LLC v. WIREdata, Inc., 350 F.3d 640, 643 (7th Cir.2003)
technically electronic, the .pdf files were not ones that anyone could readily manipulate to include on the web and link to existing real estate files. WireData brought suit against Assessment Technologies and the several municipalities in Wisconsin for failure to comply with the Open Records Act. In 2008, the Wisconsin Supreme Court gave WireData a mixed victory. The high court held that the municipalities could not deny a request for records by referring the requestor to a private third party. Furthermore, a private third party could be subject to an open records request when hired to collect data for the public entity. However, the Wisconsin Supreme Court ruled that WireData’s request for electronic records was met by the transfer of the .pdf files since they were in the requested electronic form. The high court in other words strictly construed the original request and held that it was met.

Among the lessons from the WireData is the mixed public-private partnership in the collection and dissemination of data. The municipalities, as part of their power to tax, also have the power to collect economic information that has value outside the government context. Governments outsource the information collection function to private entities that can make use of technology and business management to more effectively create databases. The problem is one of ownership and rights in the databases that are created. Open Records Acts serve to create a transparent government that allows citizens to view records created by state entities. The state cannot escape obligations of the Open Records Act through privatization. But the Act also imposed obligations on private citizens in framing requests and, under some situations, in bearing the burden and cost of disclosing the requested information. In this way, a market-like institution is created for the generation and dissemination of data.

When government data collection and open records obligations are understood as market mimicking institutions, it should not be surprising that the creation of transparent government through open records acts can also support commercial, as well as political, activities such as the ones at stake in the WireData case. The dispute illustrates the liberal perspective on data and information access, the notion that markets and politics work together to protect and ensure individual rights. Open Records Acts do not impose scrutiny on the motivation for a records request. The request could arise from mere curiosity or from the pursuit of a commercial interest, as in WireData. The government serves as a data aggregator and manager to satisfy the interests of citizens who can access the records. At the

---

167 WIREdata, supra note 163 at 755.
168 Id. at 757.
same time, the law recognizes that data protected through copyright, trade secret, or patent can be protected from request by private citizens. This private sphere within transparent government serves to protect those who process data into economically valuable products protected by intellectual property law. In this way, Open Records Acts preserve a traditional, liberal notion of markets and politics with certain data in the public sphere accessible through a political process and a proprietary sphere that supports markets. The two work together to protect individual liberty.

The scheme of data governance overlooks several critical issues. For example, under its terms, the data submitted to the government falls into the public domain unless the data is transformed into the subject of intellectual property. But this contribution to the public domain ignores issues of privacy and reputational interests that can be compromised by misuse of the data. These concerns are exacerbated by the ability of the data to be commercialized and readily used by the public. Within the liberal perspective, privacy and reputational interests may be protected by subsidiary laws, such as defamation. But such protections may not be fully adequate under an autonomy or fairness perspective. Protection of these interests would require thinking beyond individual freedom and to shape the normative framework within which market transactions and politics occur. The open source proposal in the last subsection addresses these concerns.

The first case study examined the interaction between public collection and commercial uses of data. The next case study examines the creation of a public information data library that is openly accessible. This case study illustrates similar problems with the liberal perspective, but also demonstrates potential solution through regulation.

2. Government Data

While Open Records Acts can be traced back to the Nineteenth Century as a mechanism for citizens to request information from the government, the contemporary trend, made possible by the Internet is publication of data collected by the government for ready access. This trend by governments, state and federal, to publish data has led to private entities to create search tools to identify data, aggregate it, and analyze it. Google.com, for example, provides Public Data Explorer, an application designed for journalists and researchers to analyze public databases to discover relationships and correlations in data gathered by
government agencies. Within this environment, it is not surprising that statisticians are viewed as modern day rockstars, with their ability to make data speak and use mathematical techniques as investigative tools. Against this environment, however, is the WikiLeaks controversy where the values of accountability conflict with the pragmatic concerns of security and the need for government secrecy and expertise. Government data raises a host of controversies which requires consideration of the normative foundations for the governance of data.

Adding to the controversy, public data bases serve as resources for commercialization and the development of socially useful applications. A notable example is the use of publically available information on train schedules to create a smartphone application to check on arrival and departure times for connecting trains. Software engineers explore how available data on energy usage to identify areas where conservation efforts can be targeted. These examples illustrate how private enterprise can be spurred by government investment in data collection and dissemination. Government collected data are made available to the public as resources for new applications and information based products. However, unsettled issues exist about ownership of commercially successful applications and compensation for use of government data. Although these ownership and compensation issues are far from resolved, disputes have resulted in settlements favorable to the government.

An important initiative of the Obama Administration is the creation of the website data.gov which provides access to data collected by federal agency and not otherwise protected as a government secret for security reasons. The website terms of use illustrate both a commitment by the current administration to government transparency and protection of private interests. For example, the data policy statement which is linked from the data.gov homepage limits data access that conflicts with protections for national security and privacy as defined by the Office of Management and Budget (OMB) guidelines. The statement also

---

169 Available at http://www.google.com/publicdata/home.
imposes requirements for citation to datasets downloaded from the site and obligations on the collecting government agency for data retention and data integrity.\textsuperscript{173} Finally, the statement places some restrictions on “secondary uses” of data downloaded from the site:

Data accessed through Data.gov do not, and should not, include controls over its end use. However, as the data owner or authoritative source for the data, the submitting Department or Agency must retain version control of datasets accessed. Once the data have been downloaded from the agency's site, the government cannot vouch for their quality and timeliness. Furthermore, the US Government cannot vouch for any analyses conducted with data retrieved from Data.gov.\textsuperscript{174}

Although the statement is entitled “Data Policy,” these provisions on Data.gov can be understood as terms of use which define the rights and duties of those who download from the site and make use of the data. Arguably they can be enforced by the government against individuals who violate these terms.

Particularly worth noting about the site is the distinction drawn between the concepts of “data” and “information.” In the Frequently Asked Questions (FAQ) section, the answer to the question “what is data?” is

Data are values or sets of values representing a specific concept or concepts. Data become "information" when analyzed and possibly combined with other data in order to extract meaning, and to provide context. The meaning of data can vary according to its context.\textsuperscript{175}

The distinction between data and information is critical here for a number of reasons. Under the terms of the site, public access and agency maintenance obligations apply to data, but not to information. The distinction pertains to both what users are allowed to do and how the government must carry out its duties. The term data refers to representations of some concept or set of concepts. This representation is quantitative, as the phrase “values or sets of values” suggests, but could arguably be qualitative, as perhaps the description of a certain condition of the environment. Information is made from data as the text suggests by stating that “data become ‘information’.” The implication is information is processed data, analyzed through some method, whether statistical or qualitative. Information, as

\textsuperscript{173} See supra note 12.
\textsuperscript{174} See supra note 12.
\textsuperscript{175} See supra note 12.
processed data, is not subject to the access rules nor is it subject to the limitations of secondary uses, quoted in the previous paragraph. A user under these terms can download data from the site, process it, perhaps through a statistical technique or qualitative analysis, and the end product of those analyses is information to which the user can limit access as long as the downloaded data is still made available.

The distinction between data and information is a critical one for the governance of the datasets made publically available through data.gov. The distinction allows users some proprietary protection for information generated from the publically available data while the data remains freely accessible. There is more than a hint of Lockean thinking here since information produced from the data is the fruit of the user’s labor. But there is an interpretation consistent with the values of transparency and open government. Under the terms of Data.gov, individuals are allowed to use data to form their own interpretations and evaluative opinions. These interpretations and opinions gleaned from data would be classified as information over which the user has exclusionary rights. This information becomes the tools for individual users as they engage with others in deliberation and participation in open government. What I would argue is that the terms of use under Data.gov envision an autonomous individual that can engage in the public sphere but can also retreat. This argument is consistent with the privacy protections that the terms of use permit. In short, the values reflected in Data.gov are consistent with the autonomy perspective described in the previous section.

The autonomy perspective envisions separate private and public spheres that interact with each other while maintaining independence and integrity. These separate spheres have implications for the protection of personal privacy. But they also have implications for security and government secrecy. The government also engages in the processing of data. This processing of data produces information that would not be subject to the values of open data represented in Data.gov. One example of such information is national security, a category expressly recognized within the terms of the site as exempt from open access. This exemption is included for pragmatic reasons. But the data-information distinction provides another rationale for the exclusion. National security is information that is the product of the government processing of data. It represents information that results from the autonomous sphere of governmental decision making, an analogue to the sphere of privacy enjoyed by individual citizens.

The autonomy perspective helps to explain some of the distinctions drawn by open government projects like Data.gov., such as the exclusion for national security and privacy. But the autonomy perspective fails at a basic level. If the boundaries of public and private rest on the distinction between unprocessed, or raw, data and processed, or cooked, data, how do we draw the line between processed and unprocessed? The terms of use assume that data collected by the government through its various mechanisms (tax collection, census, property information) is raw and unprocessed and does not become information until someone acts on it. But the collected data may be initially processed as well, its collection based on assumptions of how to define a unit of analysis, what questions to ask, and what tools for measurement. Furthermore, there is a question of how much processing is required for data to become information. Is simply arranging the data in some pattern enough? Does the data have to be aggregated, statistically analyzed, or summarized in some fashion? Copyright law handles these issues through the concept of originality in order to determine the boundaries of proprietary rights. It is not clear how the public and private spheres are being drawn for the purposes of data governance under the terms of the Data.gov site.

The reason for this lack of clarity is the normative thin notion of autonomy that informs data policy. Open government initiatives like Data.gov assume the zone of autonomy is defined by data processing, a sphere of cognitive decision making through which data is sorted, analyzed, and interpreted in order to aid the investigative process in which the user is engaged. The autonomy perspective leaves open the purpose behind the investigative process. It could be an aid in deciding whom to vote for, to assess the effectiveness of government policy, to identify waste and corruption, or simply to satisfy curiosity. The autonomy perspective reflected here is neutral as to goals, but seeks to provide space for the autonomous individual to define itself through engagement with data. Open government of this sort is a discussion of facts without necessarily an engagement in deeper values or commitments. There is of course a value in the discussion of facts. While such a notion of autonomy may be thin, it may also be essential for a richer development of the self. What Data.gov demonstrates, perhaps, is that the life of data may be a thin one, even if it is one that informs much of contemporary society and economy.

177 See supra note 29.
While the case study of government data may not provide a rich set of answers in its illustration of the autonomy perspective, the model of Data.gov provides a useful example that can inform the open source solution that I discuss later in this section. As I suggest, the terms of use from the web site can serve as a template for a type of norm seeding through open source licensing models for the use and dissemination of data. The open source licensing models, however, may support a richer normative framework than what is evinced by Data.gov.

3. European Databases

Open government initiatives in Europe parallel many of the developments in the United States as European governments and citizens seek to make government collected data available through web portals. Many European governments are playing catch-up with the United States through the implementation of U.S-style Freedom of Information Act legislation.179 Such initiatives demonstrate that the concerns raised in this Article about commercialization of data are global and require a comparative perspective.

Two major differences qualify some of the points raised in this Article. First, the European Union has more centralized protections for individual data privacy than the United States that treats the matter as a matter of decentralized private law rather than centralized regulation.180 As a result, many of the European initiatives for open government attempt to stave off the privacy issues raised by the commercialization of data. Second, the European Union through the 1996 European Database Directive protects database and data more strongly than the United States with the result that private incentives for the commercialization of data may be stronger.181 These two differences provide a useful case study for understanding alternative regulatory structures and their effects on data commercialization.

Professors Reichman and Samuelson, in one of the first scholarly articles on the European Database Directive, point out that one of the purposes for implementation of the directive was the promotion of database industries in Europe.182 A 2005 study of the Directive, based mostly on surveys of members of the database industry, concluded that the Directive and its implementation by


180 See Reichman & Samuelson, supra note 75 at 97-98.

181 See supra note 75 at 84-85 (describing threat to public domain from overprotection of data).

182 See supra note 75 at 72-73.
member states had little effect on the development of the industries with the United States still dominating over the European Union.\textsuperscript{183} The continuing dominance of the United States in the database industry is particularly striking because U.S. Congress had rejected the adoption of European style database protections several times during the 1990’s.\textsuperscript{184} Instead, the U.S. relies on the regulatory environment described in Section Two of this Article, a mix of weak intellectual property and common law protections. The 2005 Directive suggests that uncertainties and ambiguities over the interpretation of the rights created under the Directive are responsible for its ineffectiveness. However, the study does not address the issue of how the U.S. has managed to dominate globally despite lack of strong domestic protection for database industries.

The Directive creates a sui generis right for databases constructed through the gathering of data by the first creator. Unlike copyright law which bases protection on the originality shown in the selection, arrangement and coordination of the data, the sui generis right is established through substantial information in the gathering and collecting of data. The right extends to the prevention of the extraction of data from the database or use of the database without permission of the owner. Exceptions are created for research and educational purposes. Such protections were designed to supplement copyright protections for databases and harmonize conflicting standards for protection across members of the European Union.

The 2005 study suggests that the failure of the Directive to stimulate the database industries stems in part from the exceptions created for research and educational purposes.\textsuperscript{185} Although there is not much litigation on the scope of these exclusions, the authors of the study report that the existence of the exclusions may create uncertainties among users. Furthermore, several European Court of Justice (ECJ) opinions may have weakened the scope of the sui generis rights. Two important opinions impose limits on the exercise of the sui generis rights based on competition law principles, specifically the denial of access to a database that might constitute an essential facility.\textsuperscript{186} The two cases dealt with a television programming guide created by government run television and a database structure for recording pharmaceutical sales. The standard enunciated by the ECJ for the


\textsuperscript{184} See supra note 75.

\textsuperscript{185} See supra note 183 at 23.

imposition on a compulsory license, compelling access to data and databases raised questions about the scope of the sui generis right created by the Directive.\textsuperscript{187}

Additional uncertainty arose from the Court’s decision in the William Hill case involving access to horse racing data.\textsuperscript{188} The court ruled that an Internet database consisting of statistics pertaining to race horses was not protected by the Directive because the substantial investment in the database was not from the collection and gathering of the data, but from the creation of the data. Arguably the decision was more far reaching than the competition law decisions because the Court placed limits on what types of databases were protected by the Directive. In the William Hill decision, the Court was holding that data constructed, perhaps through some statistical or interpretative process, would not be protected by the Directive, whose goal was to protect investment in the costly gathering of data. The decision was particularly confusing because of the suggestion that the first to construct the data could not obtain protection in the database, but an entity that gathered the constructed data (perhaps by creating a database of horse racing statistics) could obtain sui generis rights in the secondary database. The 2005 study concludes that the uncertainty created by this 2002 decision may have undermined the purpose of the Directive to create stable and certain property rights in databases.\textsuperscript{189}

The experience with the European Database Directive illustrates some of the problems with establishing strong property rights in data. Underlying competition law policies limit the strength of this property right. Furthermore, there are underlying and difficult questions of what constitutes data which are unresolvable. These issues also arise in the United States context, as the cases of WireData and Data.gov illustrate. The distinction that the European Court drew between constructed and unconstructed data in the William Hill case parallels the discussion of the demarcation between data and information in the context of Data.gov. If the world of data commercialization is moving towards a regime of strong property rights and market protection, then the experience with the European Database Directive may be a harbinger of very real pitfalls to a system of strong legal entitlement.

A comparison of the prevalent approaches to data and databases in Europe and the United States suggests that a thin protection for databases and generally

\begin{footnotesize}
\begin{itemize}
\item \textsuperscript{187} See supra note 183 at 23.
\item \textsuperscript{188} Case C-203/02 British Horseracing Bd. Ltd. v. William Hill Org. Ltd., available at http://europa.eu.int/comm/internal_market/copyright/prot-databases/jurisprudence_en.htm.
\item \textsuperscript{189} See supra note 183 at 28.
\end{itemize}
\end{footnotesize}
open access to data may be more supportive of a healthy and competitive database industry. The openness permitted under the United States regime allows for competition in the provision of services that allow consumers to access and manipulate information. Competition may not be the pure textbook kind of many firms and many consumers driven by price signals. The types of competition we see in database markets may be more of the oligopolistic type with few firms competing on different margins for consumers. Nonetheless, such competition can be vibrant for an industry and may be consistent with other non-economic values such as autonomy, access to knowledge, and transparency of government and market institutions. While the first and second case studies provide some insight into different types of legal entitlement structures and their protection for values of freedom and autonomy, the third, and final, case study shows how competition informs the database industry and the market for data. The three together provide background for understanding the open source licensing model for data, presented in the next section.

B. A Proposal for Open Source Licensing of Data

How then to respond to the issues of data commercialization raised by this Article? Tracing the economics of data commercialization and the structure of legal entitlements over data and databases against a normative framework raises questions about how legal reform and legal regulation should respond. This Article has pointed to emerging problems posed by data commercialization, some traditional such as privacy and security, others novel such as the boundaries between government management of data and private uses. The case studies highlighted some of these issues with an argument presented for the role of competition as opposed to strong proprietary rights in promoting data commercialization. In that spirit, the Article concludes with a proposal for open source licensing of data.190

The proposal that follows does not address any specific problem raised by data commercialization. Instead, the goal is to provide a framework for regulation that can be used to limit some of the potential problems posed by data commercialization. The model is drawn in part from the structure of Data.gov, which provides terms of use that are meant to regulate the control and

dissemination of data provided by the site. While the specific terms were the subject of criticism in the previous section for focusing on a very narrow conception of the autonomous user, the model of terms of use is an appropriate one. Furthermore, promoting certain licensing practices has been the basis for the open source movement in software and the propagation of more user-friendly copyright terms through creative commons licenses. This Article has demonstrated that we live in a world where data commercialization will be standard practice. My proposal is to use licensing terms that can effectively regulate the practices of commercializing data. The ideas build on the work of Professor Jerome Reichman and Paul Uhlir on the scientific research commons\(^{191}\) and expands them to cover data generated in the public-private model described in this paper.

The licensing terms can be implemented in several ways. The first is to have state and federal governments that make data available through web sites or other fora to implement licensing terms that govern the use of the data. As the Data.gov example illustrates, such licensing already occurs, but unfortunately in incomplete ways. The ideas presented in this Article should serve as a starting point for identifying and addressing issues. Furthermore, these licensing terms should be propagated through private entities, such as social networking sites, that present in their terms of use the appropriate parameters for manipulating and dissemination of data generated and collected from a specific site. Competition among sites on these terms can serve as a basis for experimenting and determining which terms of use are desirable. Consideration of these terms of use should be part of good management practice for commercial sites.

What would constitute a model license for the regulation of data? I propose the following terms at the minimum:

- Attribution of the source of the data for any subsequent retransmission, manipulation or interpretation of the data;
- Limitations on reusers to limit or deny access to the data that has been gathered or collected, similar to what currently exists on Data.gov;

• Warranties that data will not be revealed or used in a way that would harm the reputation or dignity interests of those who are the subject of the data gathered or collected;
• Freedom to commercialize interpretations of the data that do not violate the warranties of reputation and dignity and the access of the underlying data;
• Freedom to replicate published interpretations of the data supported through a clear statement of methodology and approaches to the transformation or manipulation of the data.

The last model term follows practices of scientific researchers in peer reviewed journals. Its inclusion here supports the spirit of competition and openness that ensures vitality in the usage of data. The other provisions address other issues identified in this Article such as protection of autonomy interests and questions of fairness in terms of access to data and distributive values. While some people may be willing to sacrifice reputation and privacy in order to provide data that can be used for commercial gain, the model licensing terms should be designed to prevent exploitative uses of personal data.

Most importantly, these five basic model terms can support the three normative perspectives presented in this Article by allowing room for free engagement in politics and markets while protecting values of autonomy and fairness. As the methods of commercializing data evolve, the proposed open source licensing model can be adapted to address the range of concerns that develop from an adaptive marketplace. The starting point is recognizing the world of data commercialization in order to structure the regulatory response more creatively and flexibly.

V. Final Thoughts: Identifying a Legal Regime to Fit the World of Data Commercialization

Issues of data security, access, and ownership are longstanding with scholars proposing polices from the perspectives of economics, social theory, and constitutional law. This Article contributes to the discussion with the observation that the commercialization of data drives the policy debate. Consequently, the Article examines the economic and legal structure shaping data commercialization and the resulting policy concerns. Illustrating this approach, the Article concludes with an open source licensing framework for data commercialization that respects the goals of commercialization while protecting autonomy and fairness values.
This open source licensing framework as a way to continue the debate over the regulation of data collection and use which more carefully identifies the function and form of data commercialization. One point that the case studies and analysis highlight is the proliferation of data as a new resources that business and government entities can use as a source of revenue and as part of business and governance strategies. Data proliferation in both the market and political arenas highlight the potential conflicts between political transparency and commercial uses that might compromise the interests of consumers. This Article addresses these potential conflicts that elucidate the political economy of data. As data continues to proliferate, the overlap between the government and private actors, between politics and markets, demands more careful analysis of the tension between political and economic meanings of transparency. This Article is intended as a contribution to our understanding of the future of data.