Governance and Accountability in the New Data Ecology

A Vision for Electronic Data Licenses

Travis D. Breaux
Institute for Software Research
Carnegie Mellon University
Pittsburgh, Pennsylvania, United States
breaux@cs.cmu.edu

Thomas A. Alspaugh
Institute for Software Research
University of California, Irvine
Irvine, California, United States
thomas.alspaugh@acm.org

***Abstract*—Electronic data licenses (EDLs) are data governance instruments that consist of legal rules (rights, obligations and prohibitions) governing an organization's data practices. These rules include data requirements, such as rights to collect, use, retain and transfer data to third parties and prohibitions preventing these practices. We introduce the EDL concept by describing the emerging data ecology, wherein information sharing will reach unprecedented scale, and by presenting legal foundations for the EDL concept. We conclude with a broad vision for the EDL framework by discussing the license management and composition strategies, criteria for evaluating solutions, and how EDLs should support data principles and standards, before concluding with a review of related work that supports this vision.**

Keywords-data licenses, data flows, privacy and security.

I. Introduction

Information about individuals is increasingly being shared to achieve a safer, healthier, more socially connected, and more energy-efficient society. Examples in the U.S. include the Smart Grid for improved energy distribution and more intelligent energy use, various Department of Homeland Security information sharing environments for improved national security, and the National Health Information Network for sharing health information effectively and privately among healthcare providers. Examples in international commerce include Facebook, Amazon, and Twitter. Continued advances in the data sharing infrastructure and its software's design and implementation drive this increase. But the technological advances are outpacing society's progress in identifying norms, agreeing on expectations, and ensuring accountability and social responsibility. Systems are being built to use data without clear statements of the obligations of data users, assurances of the data's degree of accuracy, and provisions for recourse for those injured by careless data practices.

Current data set standards, licenses, and liability for inaccurate data are more attuned to a 1980's-era regime of primary data uses, in which consumers provided data to a product or service supplier and that supplier used it to provide the service. But now secondary uses in which data is used in a context far removed from that in which it was collected are becoming increasingly common. It is no longer the case that data users can easily assess the level of data quality as appropriate for their intended use.

Accounts have emerged in the last decade of instances in which U.S. citizens were harmed by automated use of incomplete, inconsistent, or contaminated data. These include a Kentucky woman who lost her homeowner's insurance after erroneous reports of fire damage appeared in her homeowner's insurance record maintained by ChoicePoint, an information broker; the inaccurate reports mysteriously resurfaced after repeat attempts to cleanse the contaminated data [23]. In addition, there are reports of lost job opportunities for the unemployed due to erroneous criminal record data [20] and medical prescription errors leading to harm or loss of life [4]. In each of these cases, automated or semi-automated decisions were made about individuals using data that was shared, aggregated and transformed through a network of multiple non-consumer facing parties. The individuals harmed in such cases rarely have access to this network, recourse, or assurance that the situation will be corrected. The loss of privacy, which is necessary to conduct these transactions, should be balanced with increased assurance, accountability, and social responsibility in the way data is shared and used.

Today, data transactions between online services are regulated by rules described in privacy policies, content licenses, terms and conditions, terms of service, and terms of use. For example, Figure 1 shows an example data supply chain governed by multiple rule sets: the data moves along data flows (directed edges) between different actors (circles). Each data flow is annotated by rule sets (square boxes) governing the data transaction.

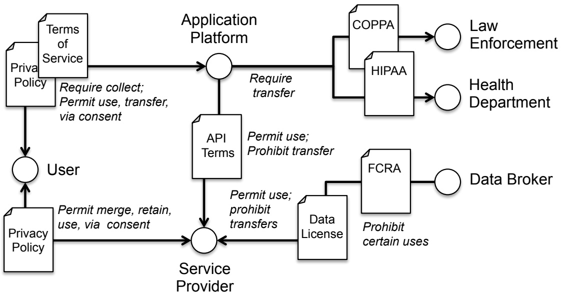

Figure 1. Example data transactions between multiple data users

In Figure 1, the user exchanges data with an application platform, such as an integrated health portal that allows users to upload medical information from doctor's visits or physical performance information from wearable exercise monitors. As an application platform, this portal will provide services (e.g., games, scrapbooks, etc.) to the user through a third-party service provider; both the platform and service provider contract with the user through their own privacy policies. The application platform's data may be subject to certain legal requirements, such as the Child Online Privacy Protection Act (COPPA), which governs information about individuals under 13 years of age, and the Health Insurance Portability and Accountability Act (HIPAA), which governs health information. In addition, the service provider purchases data from a data broker, to enrich their user's experience with new information, such as shopping preferences, location data or credit scores. Other laws, such as the Fair Credit Reporting Act (FCRA), and a separate data license between the broker and the service provider, will govern data obtained from the data broker. Each rule set contains rights, obligations and prohibitions that state how the actor is permitted, required or prohibited from using the data.

Data-governing contracts between agents in a data supply chain (including licenses, privacy policies, terms of service, and terms of use), which we collectively call data licenses, are the appropriate approach to use in regulating data practices for a number of reasons:

- Organizations are already using data licenses, although only for limited purposes and as static, text-based documents.
- The interpretation of contract licenses is relatively standardized in contract law, so the results are predictable.
- There is recognized license language (legal terminology and structures) for granting specific sets of rights and for imposing specific obligations.

There is also recognized license language for sub-licensing, so that controls established with an initial licensee can be propagated to downstream licensees. Furthermore, the rules imposed by government laws on data practices can be encapsulated and transferred within these data licenses.

However, existing data licenses are not designed so that they can be transformed and aggregrated in parallel with transformations and aggregations of the data sets they govern, producing both the output data set and the license governing its use. At present it is common that data collected for one purpose is transformed, aggregated with data collected for another purpose, and used for yet a third purpose. Data that may later be aggregated is typically collected by separate organizations acting independently, using a variety of means to collect, validate, process, and store it. Data licensors may limit or forbid specific secondary uses of their data; for example, the Experian Online Data (as of January 2011) forbids aggregation of licensed data without written consent (§3.2(a)(iv)) or use of licensed data for employment eligibility decisions (§3.2(c)(ii)). In other cases, licensors may impose additional obligations for specific secondary uses, such as obligations to obtain consent from the data subject. But the rapid evolution of data aggregation, transformations, and ultimate uses renders such relatively static measures unlikely to be sufficient. We claim that data licenses must become electronic or computational artifacts to ensure that license provisions are propagated to new data sets resulting from these transformations and that liabilities and warranties are transferred, appropriately; all in a transparent, accountable and scalable manner.

The remainder of this paper is organized as follows: in Section II, we describe recent changes in the new data ecology; in Section III, we review the legal foundations of data licenses; in Section IV, we describe the role of regulations and standards in electronic data licensing; in Section V, we propose requirements for electronic data licenses; in Section VI we present criteria for evaluating any electronic data license solutions; and in Section VII, we conclude with a review of prior work towards this vision, discussion, and a summary of the work. We use the following terms throughout this article: *data supplier* means an actor (person or organization) that maintains data and provides data to others; *data user* means an actor that uses data for a specific purpose; *data subject* means the actor about whom the data describes; and *data collector* means an actor who collects data from a data subject.

II. The New Data Ecology

The economic and social development promise of data-intensive computing is leading government and many industries to undertake unprecedented changes in data availability, data needs, and data integration. Example efforts at a national scale in the United States that stand to affect hundreds if not thousands of companies and millions of consumers include:

- In 2003, the Department of Energy promoted the Smart Grid as a means of active monitoring and control to ensure "a two-way flow of electricity and information between the power plant and the appliance" [22].
- In 2005, policy advocates acknowledged the transformative effects of using the planned Nationwide Health Information Network (NHIN) for conducting medical research across millions of medical records, nationwide [29].
- In 2009, the White House launched the Data.gov website to make "economic, healthcare, environmental, and other government information available on a single website" [25].

In the private sector worldwide, Internet "mash-ups" and mobile applications on smartphones integrate data from disparate sources to produce unprecedented tools for personal decision-making. Apps provide services in support of decisions involving health, finances, route-finding, weather, and detailed information about consumer products from their barcodes. The data and decisions range from confidential to non-critical, and consumers have accepted occasional inconvenience or waste due to inaccurate data as the cost of using these novel services, perhaps because they are still novel or because the individual costs of inaccuracies seem small. But as apps and mash-ups are integrated into personal and business practices and become increasingly critical, consumers and organizations will need stronger assurances based on data controls and selective transparency of data sources and practices.

In contrast, automated and semi-automated decision-making in highly integrated data supply chains relies implicitly on high quality data. A key data licensing issue here is propagation of warranties of data quality and liability for damages due to inaccurate data. The importance of warranty and liability provisions varies depending on the necessary level of confidence in data and the availability of emerging techniques to detect and correct data errors; where the consequences of inaccuracy are catastrophic or can potentially be multiplied by downstream data systems, the need for them is great. An inaccuracy that appears innocuous in one context can result in harassment, job loss, financial loss, damage to health, or loss of life in another context [20, 23, 4]. As data sources and error detection and correction improve, on the other hand, data suppliers may choose to offer warranties and assume some degree of liability to gain competitive advantage.

Automated decision-making produces a decision using input data from a range of sources; the consequences of the decision may be far greater than expected for the character of any of the individual input data sets. It is not reasonable for the organization automating the decision to bear all liability for an inappropriate result when the input data is inaccurate, nor for the price of data whose accuracy may have large consequences to be the same as that of data whose accuracy is less significant. The liability will need to be shared and transmitted back down the data supply chain to the supplier of inaccurate data, and licensing is the most efficient, flexible, and appropriate means to do this. Market mechanisms can then help set appropriate data prices and drive the adoption of appropriately responsible data practices, over the wide range of data collection contexts and emerging data applications. We believe mechanisms such as these can mediate the use of data of mixed quality and of incomplete data.

III. Legal foundations for licenses

Under U.S. and other national laws, a license grants one or more rights to property (intellectual or otherwise) without granting ownership. These rights may include the right to sublicense, by which the licensee can in turn license certain rights in the same property to others. A license is a *bare license*, if it simply grants rights, or a *contract license*, if it grants rights in exchange for obligations or other consideration. The consideration given by the licensee can be small, as little as a peppercorn in the traditional explanation, but it cannot be nothing. In addition to a payment or other such consideration, it can take the form of: (a) an act other than a promise; (b) a forbearance; or (c) the creation, modification, or destruction of a legal relation. Because contract law is more settled in the United States and other countries than the law of licenses, most software and data licenses are contract licenses. Contract licenses are flexible in the range of activities they govern and dynamic in their ability to direct legal obligations along paths determined by how the licenses material is used by counter-parties [28].

In order for a license to have a useful effect, the licensor must possess rights that are not freely and generally available to everyone. In the case of intangible goods such as data, the basis for ownership is rooted in the law of intellectual property, primarily copyright. Copyright law gives to the creator of an original work certain exclusive rights in that work, such as the right to reproduce the work, prepare derivative works, distribute copies, and sublicense any of these rights, including the right to sublicense, or parts of those rights to other parties. In the United States, copyright is regulated by the U.S. Copyright Act (17 U.S.C.), and in most of the world it is regulated by the Berne Convention for the Protection of Literary and Artistic Works.

Contract licenses may contain provisions for other purposes, as long as both parties are willing to agree to the provisions. For example, in some software licenses the licensor indemnifies the licensee against certain kinds of liabilities, such as certain kinds of infringements of patents controlled by the licensor; an example is the Apache License, version 2.0. The obligations imposed on licensees vary by license, as do the provisions by which rights and obligations are transferred in sub-licensing. Contract licenses may connect the fulfillment of obligations to the rights granted, in order to achieve goals beyond copyright protection. For example, open source software (OSS) licenses use this mechanism to increase the commons of open source software available to all, to ensure attribution rights for open source developers and to drive the development of open source communities and ecosystems [28]. Whereas the aim of OSS licensing is often to make source code broadly available, we expect that data licensing has other aims, such as ensuring data is of high quality or ensuring security and privacy.

Contract licenses may disclaim certain standard obligations in regular business transactions, such as warranty. For example, there are implied warranties of merchantability, wherein the licensor implicitly guarantees that goods conform to reasonable expectations and have no hidden faults, and implied warranties of fitness for a particular purpose, wherein goods are guaranteed to be suitable for their stated purpose. Warranty in the United States is not covered by federal law, but rather by state law. All U.S. states adopted the Uniform Commercial Code (UCC), and some states also adopted other codes such as the

Uniform Computer Information Transactions Act (UCITA). However, the UCC is not interpreted uniformly nationwide, because states adopted different revisions of the UCC, and courts in different states may interpret a provision in different ways. In addition, states have legislated warranty and liability protections beyond those in the UCC.

Software and data licenses frequently attempt to disclaim warranties and liability. However, this extreme is deemed unethical in some domains, such as healthcare [12], and indeed one goal of the UCC is to set up mechanisms for equitably allocating the warranty and liability obligations of sellers, resellers, and buyers. We believe electronic data licenses can use finer-grained warranty and liability provisions that scale with different levels of data quality. Warranty provisions can establish the level of accuracy warranted by data suppliers. Liability provisions can require equitably distributing potential damages from unreliable data among all licensors and licensees involved in a transaction. These provisions provide licensor and licensee with a mechanism for converging on a price that accounts for the warranted value and the cost of liabilities. Finally, sub-licensing provisions may restrict rights to limit data uses and delineate the context and scope in which warranty and liability provisions should be interpreted.

IV. Regulatory Sources of Data Requirements

United States and international laws and standards increasingly govern data practices. These laws describe contract mechanisms to ensure data users implement specific practices, including privacy and security safeguards, as well as describing controls on specific data types. In addition, industry may self-regulate by introducing their own industry data standards.

A. Legally required data practices through contracts

National and state laws impose data requirements on third-party data users through business contracts. For example, the U.S. Health Insurance Portability and Accountability Act (HIPAA) Security and Privacy Rules enacted in 2003 require covered entities, including hospitals, insurance issuers, and health plans, to document in contract language that their business associates will implement the privacy and security requirements of the Rules[1] and that these requirements will be propagated via other contracts to any agent or subcontractor of the business associate who handles protected health information.[2] In addition, the Gramm-Leach-Bliley Act (GLBA) Safeguards Rule requires financial institutions to require service providers to implement and maintain reasonable safeguards for data.[3]

Some laws require statements describing "how" service providers will safeguard information, whereas other laws specify the types of statements, in which case the contract language varies greatly by institution. While some of these contracts are negotiated during the establishment of the business relationship, the contract language can also appear in a data license that is bound to the data. We believe this last approach will become more prominent, especially as business relationships are built around sharing heterogeneous data sets that are governed by different data standards and legal requirements dependent upon the type of data or business practice, as we now discuss.

B. Legal limitations on data types and data practices

During the past decade, forty-six U.S. state data breach notification laws were passed by state legislatures placing strict requirements on the handling of personal information of residents of each state. While these laws aim to encourage security best practices, they also contain conflicts and inconsistencies that make it more difficult for businesses and institutions to exchange and manage data in a uniform and consistent manner. Such conflicts concern what kind of data or for what purpose the data is used and what steps are taken to protect the data. For example, a data user who receives personal information on a Massachusetts or Maryland state resident is bound to notify the resident of any breach of security affecting that data. However, under Massachusetts law §93H-3, the notice must not include the nature of the breach, whereas under Maryland law §14-3504(g)(1), the notice must include a description of the breach. In addition, these laws cover different types of information. For example, Maryland law §14-3501(2)(iii) does not cover health information governed by the HIPAA, whereas Arkansas law §4-110-103(7)(D) specifically covers medical information. Finally, laws have different standards of care associated with required data practices: Massachusetts law 201 CMR 17.02 requires encryption using a confidential process or key, whereas Nevada law Ch. 603a requires standards-based encryption that has the potential to be higher quality.

U.S. federal laws also limit the scope for which certain types of data may be used. The Fair Credit Reporting Act (FCRA) regulates information qualifying as part of a consumer report by restricting the disclosure of this information on a purpose-based, positive control list, including purposes for obtaining employment, insurance and banking credit. The HIPAA Privacy Rule places restrictions prohibiting the disclosure of DNA and dental records to law enforcement agencies.[4]

U.S. state laws have varying degrees of restrictions on types of health data. The California Health and Safety Code §121025(a) prohibits the disclosure of HIV test results, except as required by law or with the written authorization of the patient. In contrast, the New Mexico Health and Safety Code §24-2B-6(A) permits the disclosure of HIV test results to healthcare employees with "a need to know," introducing a potential conflict

[1] 45 CFR 164, §§308(b), 502(e)(1)

[2] 45 CFR 164 §§314(a)(2)(i)(B), §504(e)(2)(ii)(D)

[3] 16 CFR 314 §4(d)(2)

[4] 45 CFR 164 §512(f)(2)(ii)

with the California Code. The challenge for existing and future information systems is the ability to rationalize these complex, multi-jurisdictional requirements in a manner that can be verified in a transparent, accountable and trustworthy manner. Due to the sheer scale in number of requirements, our position is that this challenge can only be addressed computationally to scale with the number of new and innovative applications afforded by current and future technology. Furthermore, we believe these requirements must "follow the data" to ensure that data shared for secondary purposes continue to afford its original protections.

C. *Data structure and management standards*

Data practices have evolved in profound and life-changing ways over the past three decades. When a data practice is new, it may not be regulated by government, and may introduce privacy and security risks. To preempt regulation and reduce market risks, industries often develop data standards describing how an organization should structure and manage data. Data structure standards, such as the Health Level 7 (HL7) messaging standards in healthcare, provide insight into the types of data elements exchanged, for what purpose, and with whom. These standards are useful for understanding data flows in which transparency in business-to-business data exchanges may otherwise not exist. Furthermore, data licenses may require the use of standards to improve reliability in data exchange and to place requirements on specific data elements using a shared data vocabulary.

Data management standards, such as the Payment Card Industry (PCI) Data Security Standard (DSS) that governs retail payment information, include a mix of administrative or policy requirements and technical software requirements. These requirements can affect data exchange in ways that are similar to laws, for example, by requiring companies to implement network security best practices, such as network segmentation to reduce the risk of data breaches. Standards can also serve as powerful proxies in compliance with laws: for example, Nevada law Ch. 603a.215(1) requires data collectors who accept payment cards to comply with the current version of PCI DSS. Furthermore, PCI DSS requires businesses that use third parties in payment processing to propagate requirements to the third party. We believe that data licenses can offer new, transparent, and accountable mechanisms for distributing administrative and technical obligations found in data management standards across data supply chains. By allowing the obligations to follow the data in a computationally verifiable manner, we believe businesses will gain greater flexibility in choosing sub-contractors to reach greater market efficiencies.

V. ELECTRONIC DATA LICENSE FRAMEWORK

Electronic data licenses (EDLs) provide data suppliers and data users with a common framework to share and use data while managing traceability between data and governing requirements in contracts and laws. The framework defines a *data context* as a bounded set of data practices. A data context may correspond to an organization's entire set of data practices or to an individual business unit, thus providing greater separation among data practices and further definition of data transfers between business units. In addition, data contexts could be defined to bound practices governed by a specific set of laws, such as COPPA or the FCRA. Thus, data contexts may overlap where business units overlap with multiple regulations.

Data contexts consist of a set of inputs, or data collections, and a set of outputs, or data transfers. Activities that take place within a context include data aggregation, analysis, use, retention, etc. The choice in granularity for an organization (one context or multiple, interconnected sub-contexts) depends on organizational size and need to fine-tune their data practices to avoid unwanted uses and disclosures. By treating data contexts as "buckets" of practices, as opposed to explicit inter-linked workflows with pre- and post-conditions as seen in business process modeling[5], organizations have greater flexibility to evolve their practices within these contexts without the burden of redrawing workflows in their models.

The "bucket" viewpoint fits with the assumption that centrally or hierarchically managed organizations still have limited insight into their business units. Evidence for this assumption in the U.S. exists in the GLBA, which requires CEO's to certify their business practices comply with the law, and FTC enforcement actions finding that business practices are misaligned with privacy policies [8]. For example, an employee can envision a new way to use data by extending an existing or inventing a new business practice. This employee can declare the "use" of this data within their data context and then use EDLs to determine how to acquire the data, if its not already present in their context, and what rights and obligations govern that data. New rights can be requested, if they do not exist. Apart from declaring the data purpose, the employee does not have to detail *how the data will be used* or *with whom the data will be shared* within their context, although these declarations may be required by the organization for other reasons.

To better understand the framework and data context, consider Figure 2, which illustrates data flows in two data contexts: data actions (circles) are connected by data flows (directed edges). Each action is labeled by a verb describing the operation performed by the actor on the data as it moves along the data flow:

- COLLECT – data is collected by a data collector from a data subject
- CREATE – new data is created from existing data, for example, by calculating a person's shopping preferences from their shopping history

[5] See the Business Process Modeling Notation (BPMN) at http://www.bpmn.org

- TRANSFER – data is transferred from a data supplier to a data collector or data user
- RETAIN – data is retained for a period of time in a data store, which allows reuse as new data flows are created from this data store
- USE – data is used by a data user to perform a business practice
- MERGE – data is merged from multiple sources to create an aggregate dataset

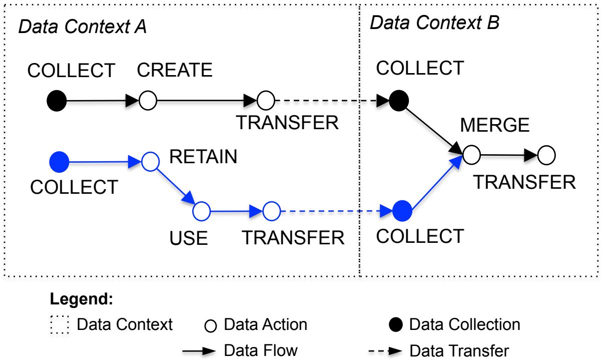

Figure 2. Example illustrating data flows within and across data contexts

In Figure 2, there are two data flows in data context A: the top flow leads to a creation of new data, which is transferred to context B; the bottom flow (in blue) leads to a retention, use and onward transfer to context B. In context B, the data is collected from context A, merged into an aggregate data set and transferred onward. These two contexts represent only a fraction of the data flows that can perceivably exist within an organization. Data flows may be further subdivided to distinguish data user, data subject, data object, data purpose. EDLs contain expressions that permit, require or prohibit data actions in these flows, but do not specify the specific flows: for example, by prohibiting the aggregation or creation of new data from existing data, or by preventing onward transfers without consent of the data subject.

The data actions define a traceability matrix, which allows two-way communication of events, such as new data requirements from data collection points to data transfer points, and vice versa. In Figure 2, for example, the data user in data context B can propagate messages "upstream" to data context A, who in turn could propagate these messages further upstream through their own data collection points. Such real-time messaging can help data users share privacy and security threat information, such as data breach notices, and license changes, such as new rights or prohibitions due to data subjects invoking or revoking their consent.

With this example in Figure 2 in mind, we now discuss two contributions of the framework: license management and license composition. These two contributions were gleaned by visually inspecting the following data licenses:

- Facebook Platform Policies – Facebook Developers (12/22/2010)
- Facebook Privacy Policy (12/22/2010)
- Google Health API - Terms and Conditions (4/24/2009)
- Google Health Developer Policies (9/2/2010)
- Google Health Privacy Policy (9/2010)
- Google Health Terms of Service (4/28/08)
- Experian Online Data License Terms and Conditions (11/12/2010)

The visual inspection consists of coding individual statements in the license contract to determine what impact each one has on a possible framework. Because these inspections were exploratory and sought to identify necessary requirements for developing a framework, the inspection is not a complete analysis. Thus, the proposed framework that follows may need to be extended to support the full range of data license expressions.

A. License Management

License management includes activities required to maintain licenses during the data lifetime as follows:

- **Activation**: Passive or active events that trigger specific clauses, such as "By visiting this website, you agree to…" and "To use this service, you must…."
- **Termination**: Conditions under which the license may be terminated.
- **Notification**: Conditions under which a licensor or licensee may be notified of events, including license violations, changes to the data or changes to the license.

To effectively manage EDLs, data users must have open, two-way communication channels with data suppliers. Each channel is opened when an EDL is activated, for example, by creating an remote access account to receive data via a web service; the channel is closed when the license is terminated and the data is either released from the license or destroyed, per the license agreement. While the channel is open, a data supplier can post license changes to the data user, which may result from new regulatory rules or treaties, for example. These changes may only affect new data transfers or the changes may "grandfather" in old data previously collected from the data supplier.

In addition, the data user can post notices to the data supplier, which may include pre-defined events triggered while in custody of the data, such as data breaches or data quality concerns. The scope and content of these messages is beyond the focus of this paper, however, we recognize the need to broadly describe this contribution of the framework, since it is necessary to maintain licenses over time.

Communication between data subjects and data collectors can also use these channels. When a data subject visits a website, the channel is opened with rights and obligations assigned by the subject to the website in conformance to the website's privacy policy; effectively, giving the website a license to collect data about the subject. When the subject creates an account, they may further modify this license by assigning or revoking rights via an opt-in/opt-out consent mechanism about sharing their data with third parties. If a website or third party wishes to repurpose data, e.g., use data for a new purpose beyond those for which the data was originally collected, the new rights can be communicated via this channel to the data subject, effectively notifying the data subject about the new rights.

The edge set for a data context (see Figure 2), which consists of the directed edges between data actions, can be inferred from a data context, if the licenses explicitly define from who data is collected, the data type, the data purpose and to whom data is transferred. As previously mentioned, there may be several different business practices performed under the scope of these rights and obligations, as long as these practices fall under the stated data purpose. This edge set yields a traceability matrix that can be used to trace data from any collection action to any transfer action. These traces are used to send messages across data contexts, thus linking communication channels in the data supply chain and enabling the downstream and upstream propagation of notices and data requirements.

Today, these license management activities are either unrecognized by stakeholders or performed in an ad-hoc manner. The responsibility to detect changes in privacy policies, terms-of-use and terms-of-service agreements largely falls on the data subject or data user, who must revisit these online documents for manual inspection. As distributed systems continue to scale up in size and increasingly perform automated decision making, these ad-hoc processes can yield increased system failure and become unmanageable by users. MacDonald and Cranor calculate that it would take an average person 244 hours per year to read all the privacy policies for websites they visit, excluding changes to privacy policies [18]. Thus, stakeholders need a coordinated framework to publish, trace and account for changes across data supply chains.

B. License Composition

Electronic data licenses are composed from rules, such as rights, obligations and prohibitions, governing data practices. We now describe the types of activities governed by an EDL, the source of authority to regulate these activities, and how EDLs can minimally formalize the meaning of compliance with a license.

Based on our analysis of text-based data licenses, we identified the following categories of rule-governed activity; these categories may not be complete, however, we found them to be *necessary* to accommodate the licenses that we inspected:

- **Alteration**: Conditions under which data can be altered, modified or derived.
- **Assignment**: Conditions under which rights or obligations may be re-assigned to other parties.
- **Coverage**: The types of actors, data and data purposes covered or not covered by the license.
- **Distribution**: Conditions under which data can or cannot be used or shared.
- **Ownership**: Conditions under which data is owned and the identity of the owner.
- **Quality**: Conditions describing the data quality and content assurances, or lack thereof.

We envision organizations using EDLs as data governance instrument to create and propagate rules governing data practices. A *power* is the ability to assign rights, obligations, and prohibitions to other parties [13] and the rules in an EDL originate from two separate powers: (1) the power of governments to regulate data, industry-wide; and (2) the power of data owners or data stewards to regulate data that they possess for their own personal or business reasons. Using the second power, data suppliers can transfer their powers to data collectors, or assign specific rights, obligations and prohibitions to data collectors. Thus, an EDL contains rules that originate from either one of these two classes of legal power and this origin must be distinguished in the license composition.

Rights, obligations, and prohibitions have a well established foundation in law [13], have been observed in numerous case studies of formalizing privacy policy and regulations [9, 6, 7] and can be expressed in Deontic Logic [15] using the following axioms to detect conflicts:

```
A1: Obligation(x) → Right(x)
A2: Prohibition(x) → ¬Right(x)
```

Axiom A1 states, if an actor is required to perform the action *x*, then that actor is also permitted to perform the action *x*; thus, obligations imply rights. Axiom A2 states, if an actor is prohibited from performing the action *x*, then that actor is not permitted to perform the action *x*. Using Deontic Logic, we can detect conflicts between rights, obligations, and prohibitions. Thus, we say a set of data practices P complies with a license L, if and only if, $L \models P$, which means every practice in P is at least permitted by one right and not prohibited by any prohibitions in L. This definition is also supported by the least-fixed point logic [11].

Exclusions describe an act that is not expressly permitted, required or prohibited by a policy [5]. Exclusions can be used to enforce gaps in data licenses, which are then “filled” by other licenses, for example, when two data sets governed by separate licenses are merged and the first license contains rules that fill a gap in the second license. Another mechanism for creating gaps in licenses is the use of exceptions when declaring classes of actor, data or purpose covered by the license. For example, a license may exclude a person’s street address from their mailing address, leaving only the city, state and zip code in a permitted transfer. A separate type of exception can establish priorities between rules [6].

In addition, in licenses we observe new notions of warranty (assurance that facts are true or events will occur), liability (being responsible for events) and indemnity (security or protection against losses) that must be formally coded and accounted for. Warranty, liability, and indemnity provisions are rights and obligations that take effect under specific conditions, when an error in data or other fault of a data supplier damages a downstream data user or when a data user makes inappropriate or prohibited use of supplied data. In such cases legal liability, financial damages, and restitution follow the paths recorded in the traceability matrix. Such provisions have the potential to play a significant role in industry self-regulation.

VI. Criteria for Evaluating Solutions

We envision several criteria for evaluating research in this area, including the ability to formally check consisting within and among EDLs, support for operations on EDLs in parallel with operations on data, and support for producing natural language formulations from the result transformations that have appropriate legal effects. To this end, EDLs may be realized in many ways, including by combining existing technologies into a novel architecture. We now describe a minimum set of evaluation criteria for vetting prospective solutions.

Expressiveness. EDLs are expressed in a machine-readable language with a formal semantics. We identified the following core language features for expressing an EDL:

- *Prescriptions* – statements that are annotated as a right, obligation, or prohibition. Some prescriptions may be provisional, or conditioned on another prescription, e.g., *if* one acts, *then* one must…
- *Powers* – statements to assign prescriptions to other actors; powers require maintaining provenance between prescriptions and their originating authority
- *Actor and purpose classes* – a reusable classification hierarchy for describing actors and purposes
- *Data classes* – a reusable classification hierarchy for describing data elements with compositionality, or the additional ability to describe data elements as components of data sets and lists
- *Extensible actions* – the minimum actions for alterations (e.g., merge, create) and distributions (e.g., collect, transfer, retain), and the ability to define additional actions as data practices evolve

These core language features are well-established concepts and appear in many authorization languages, such as P3P [10], E-P3P [1], EPAL [26], XACML [21], and Rei [17], and usage control models, such as UCON$_{ABC}$ [26] and the Data -Purpose Algebra [14]. However, these languages do not address the following issue of compliance and transformability.

Compliance and Complexity. EDLs provide a means to logically verify compliance using automated tools. We envision compliance verification as a reasoning task in formal languages, such as first-order logic or graphs. Entailment is a reasoning task in propositional logics in which a set of premises P entails a conclusion c, written $P \vDash c$, if and only if, every interpretation that satisfies the premises also satisfies the conclusion [16]. In EDLs, this task may be used to show that a property of a license, such as a right to use a class of information, is true for all interpretations in that license. Thus, if a prohibition conflicts with this right, the entailment task would yield a contradiction, or an interpretation where the right is false. To check entailment, one can examine a truth table, which consists of generating all 2^n interpretations for n symbols. For licenses with a large number of symbols to describe the prescriptions in the license, this naïve method is computationally infeasible. Thus, any solution to EDLs must show the complexity bounds of the reasoning method in the logic as a function of the number of symbols expressible in the language.

We define three compliance verification goals:

1. *Internal Consistency* – a license is free of internal conflicts. The license L is internally consistent, if and only if, for all permitted actions in L, there exist no prohibitions over those actions in L. To verify a compliance test and external compatibility, we assume that licenses are first internally consistent.
2. *Compliance Test* – an atomic test that is used to determine if an existing or envisioned practice is permissible, required, or prohibited under a license L. For example, we might check that the practice p complies with a license L by showing that $L \vDash p$, which is true if at least one right permits the practice and no prohibitions prohibit the practice.
3. *External Compatibility* – a license is compatible with another license, if the set of permitted, required and prohibited actions in the first license are contained in the second license. The license A is compatible with the license B, if and only if, for every expression $a \in A$, there is an expression $b \in B$ such that a is subsumed by b. For example, if a and b are rights, we say that b is at least as permissive as a, and a is no more more permissive than b.

Other verification goals can be envisioned, however, the above goals provide basic mechanics for showing that a license is conflict-free, for allowing data users in large organizations or across organizational boundaries to query licenses for existing or envisioned practices, and for determining whether two licenses conflict. External compatibility can also be used to check whether a set of practices complies with a license by generating a set of rights, wherein each practice corresponds to a single right, then checking whether this set of rights is compatible with the license.

Transformability. EDLs exhibit the property of transformability, in which each data operation that is expressible in an EDL and performed on covered data elements has a corresponding license operation that is performed on the governing license(s). We consider two data operations:

The *create* action accepts one or more data elements as input and produces a new data element as output. After creation, what set of prescriptions govern the new data? One approach permits license authors to write sublicenses for specific actions, such as create. If the data user creates new data, then the new data receives the sublicense: a separate set of explicit prescriptions from the original license. Alternatively, if no sublicense is specified, the default assumption may be to confer all obligations on the input data element(s) to the new data element, and all rights for the input data element(s), or their intersection if two or more input EDLs are involved. Thus, licenses may permit creation, but prohibit onward transfer of the new data, for example.

The *merge* action accepts two or more data elements as input and yields a new dataset containing the input data elements. Such aggregation of data yields a new license that covers the output dataset. This new license could consist of: (1) the intersection of the rights, such that the new license is no more permissive than any of the input licenses; and (2) the union of the obligations, such that the new license is no less obligatory than any of the input licenses. The rights and obligations combine in the same pattern seen in the virtual license for a software system composed of heterogeneously-licensed components [2, 3]. This license operation guarantees that there is *no gain in new rights* to the data and there is *no loss of existing obligations* on the data.

Other Principles and Standards. Formal languages for expressing EDLs should support the assignment and verification of principles and standards to data contexts.

Anonymity and its converse, identifiability, have increasingly been discussed as a privacy risk in the new data ecology [24]. It is assumed that data in some combination is identifiable [30] and that adding new data to existing datasets is a method for re-identification [19]. In cases where these combinations are well known, an anonymity principle may prohibit merging data to yield an individual's identity from the combined dataset. Alternatively, a licensor may wish to declare that certain data elements, such as a bank account holder's name and account number, are encrypted when contained in the same dataset.[6]

In addition, the OECD Guidelines on the Protection of Privacy and Transborder Data Flows describe the collection and use limitation principles, which state that data will only be collected when a need has been expressed, or that data will only be used for the purpose for which it was originally collected, respectively. Similarly, the HIPAA Privacy Rule contains a data minimization principle, which restricts the transfer of data to only that which is necessary for a specific purpose. The use limitation principle can be formalized in first-order logic, when data purposes are specified [14], and we envision that the collection limitation or data minimization principles can also be formalized. The ability to guarantee these principles across data flows (and derived licenses) is a compelling advantage of EDLs and any languages that fulfill this vision.

VII. Discussion and Summary

The challenge posed by electronic data licenses is new and, to our knowledge, has not been addressed by prior work. Multiple languages exist to date that support the expressive characteristics of EDLs, which includes both authorization languages (XACML, EPAL, E-P3P, Rei) and usage control models ($UCON_{ABC}$, Data-purpose Algebra). With the exception of Rei, which is a superset of OWL-Full and thus undecidable, the other languages require additional validation to demonstrate the remaining EDL features we describe in Section V. Principally, this validation must demonstrate support for the transformability principle and the ability to compare licenses with reasonable complexity bounds for real-time verification. Reference implementations, which are used to validate and demonstrate the language capabilities, are desirable for industry and research use.

We contrast the expressive characteristics needed for EDLs with the simpler model used in our prior work on open source and proprietary software licenses [2, 3]. That work expressed software licenses in terms of rights and obligations, only, based on Hohfeld's concepts of right, privilege, duty, and no-right [13]. The goal of this work is to provide automated guidance for software designers and integrators to ensure that systems built from heterogeneously-licensed components could legally be used, distributed, and evolved, in exchange for acceptable license obligations. In this context, the only stakeholders were the licensors and licensees, who were assumed to be acting in good faith. The kinds of interactions among licenses were controlled by system fundamentals such as the concepts of compiling, linking, and distribution, which evolve but at a measured pace. The laws governing software licensing are stable and well established. The tight focus of this goal made it possible to characterize more-complex license provisions, such as termination of licenses, if the licensee modifies a library in violation of certain conditions (e.g. Lesser General Public License version 2.1 §8) or makes a patent infringement claim against the licensor (Academic Free License v. 3.0, §10), as simple prohibitions. In contrast, the data license context involves a larger variety of stakeholders and different patterns of licensor-licensee interactions. The goals of EDLs are more complex, novel, and still under exploration, so it remains to be seen what specific classes of license provisions are irrelevant. The potential kinds of interactions among data licenses, such as the types of licensee, data governed and purposes of data use, are more varied and rapidly evolving. Unlike software integration, data use is a highly regulated area as discussed in Section IV. As a result, the necessary expressive characteristics of EDLs require a more complex formal foundation.

Aside from the theoretical challenges and evaluation criteria, there are also industrial challenges to realize this vision. In this paper, we particularly focused on how EDLs express data requirements. In practice, data is

[6] This encryption principle is desirable to comply with several U.S. state data breach notification laws.

handled by thousands of different technologies and products. Data transfers, for example, can occur over networks using a wide variety of communication languages and protocols, or across physical spaces using data storage devices. For practical purposes, we envision at least two views for aligning EDLs with these technologies. The traditional view treats EDLs as a set of requirements that are aligned with product designs prior to and during development and integration of these technologies to "prove" the technology respects the prescriptions in the license. However, modern computing systems are increasing dynamic: configurations can change at runtime, such as the dynamic loading of plugins or rewriting of rules that govern system behavior. This dynamicity requires a second view, wherein EDLs and reconfigurable systems are validated during runtime to check that the new behavior conforms to the license. Designers who incorporate dynamicity into their software using abstractions, such as plugins, would need an interface in the abstraction to verify that the plugin does not violate the EDL. Furthermore, this verification must support plugins not yet envisioned but still supported by the system. We acknowledge the magnitude of this challenge and that future work must investigate appropriate boundaries for employing EDLs in the presence of robust dynamicity.

ACKNOWLEDGMENTS

This material is based upon work supported by the U.S. Air Force under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

REFERENCES

[1] P. Ashley, S. Hada, G. Karjoth, and M. Schunter, "E-P3P Privacy Policies and Privacy Authorization," In Proc. *ACM Workshop on Privacy in the Elec. Society*, 2002, pp. 103-109.

[2] T. A. Alspaugh, H. U. Asuncion, and W. Scacchi. The role of software licenses in open architecture ecosystems. *1st Int'l Work. on Soft. Eco. (IWSECO-2009)*, pp. 4–18, Sep. 2009.

[3] T. A. Alspaugh, W. Scacchi, and H. U. Asuncion. Software licenses in context: The challenge of heterogeneously-licensed systems. *J. Assoc. for Info. Sys.*, 11(11):730–755, Nov. 2010.

[4] P. Aspden, J. Wolcott, J.L. Bootman, L.R. Cronenwett, "Preventing Medication Errors: Quality Chasm Series," Committee on Identifying and Preventing Medication Errors, National Academies Press, 2007.

[5] T.D. Breaux, "Legal Requirements Acquisition for the Specification of Legally Compliant Information Systems," Ph.D. Thesis, North Carolina State Univ., Raleigh, NC, May 2009.

[6] T.D. Breaux, A.I. Antón, "Analyzing Regulatory Rules for Privacy and Security Requirements," *IEEE Trans. Soft. Engr.,* 34(1):5-20, Jan./Feb. 2008

[7] T.D. Breaux, A.I. Antón, Kent Boucher, Merlin Dorfman, "Legal Requirements, Compliance and Practice: An Industry Case Study in Accessibility," In Proc. *IEEE 16th Int'l Req'ts Engr. Conf.*, Barcelona, Spain, pp. 43-52, Sep. 2008

[8] T.D. Breaux, D.L. Baumer, "Legally 'Reasonable' Security Requirements: A 10-year FTC Retrospective", *Computers and Security*, 30(4): 178-193, 2010.

[9] T.D. Breaux, M.W. Vail, A.I. Antón, "Towards Compliance: Extracting Rights and Obligations to Align Requirements with Regulations," *14th IEEE Int'l Req'ts Engr. Conf.*, Minneapolis, Minnesota, pp. 49-58, Sep. 2006.

[10] L. Cranor, M. Langheinrich, and M. Marchiori. The Platform for Privacy Preferences 1.0 (P3P1.0) Specification: W3C Recommendation, 16 April 2002. http://www.w3.org/TR/P3P/

[11] H. DeYoung, D. Garg, L. Jia, D. Kaynar, A. Datta, "Experiences in the logical specification of the HIPAA and GLBA privacy laws," *ACM Work. on Privacy and Elec. Soc.*, 2010, pp. 73-82.

[12] J. Goedert, "AMIA: Hold Harmless Clauses Unethical," *Health Data Management*, 11 Nov. 2010.

[13] W.N. Hohfeld. "Some fundamental legal conceptions as applied in judicial reasoning," *The Yale Law Journal*, 23(1):16–59, 1913.

[14] C. Hanson, T. Berners-Lee, L. Kagal, G.J. Sussman, D. Weitzner, "Data-purpose algebra: modeling data usage policies," 8th IEEE Work. Pol. Dist. Sys. & Nets., 2007, pp. 173-177.

[15] J.F. Horty. "Deontic logic as founded in non-monotonic logic." *Annals of Mathematics and A. I.*, 9: 69-91, 1993.

[16] M. Huth, M. Ryan. *Logic in Computer Science, 2 ed.* Cambridge University Press, 2004.

[17] L. Kagal, *A Policy-Based Approach to Governing Autonomous Behavior in Distributed Environments*, Ph.D. Thesis, University of Maryland, Baltimore County, Sep. 2004.

[18] A.M. McDonald,L.F. Cranor. "The cost of reading privacy policies*," I/S: A Journal of Law and Policy for the Information Society. 2008 Privacy Year in Review issue.*

[19] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, Muthuramakrishnan Venkitasubramaniam. "L-diversity: privacy beyond k-anonymity." ACM Transactions on Knowledge Discovery from Data, 1(1): 3, 2007.

[20] S. Michels, "Advocates complain of background check errors: dozens of lawsuits claim lost jobs from inaccurate criminal records", ABC News, 13 Oct. 2008.

[21] T. Moses, ed. eXtensible Access Control Markup Language (XACML) Version 2.0, Oasis Standard, 1 February 2005. http://docs.oasis-open.org/xacml/2.0/

[22] Office of Electric Transmission and Distribution, "Grid 2030: A National Vision for Electricity's Second 100 Years," United States Department of Energy, July 2003.

[23] R. O'Harrow, *No Place to Hide*, Free Press, 2006.

[24] P. Ohm, "Broken promises of privacy: responding to the surprising failure of anonymization," UCLA Law Review, 57(6): 1701-1777, 2010.

[25] P. Orszag, "Democratizing Data", White House Press Release, 21 May 2009.

[26] J. Park, R. Sandhu, "The UCONABC usage control model," *ACM Trans. on Info. and Sys. Sec.*, 7(1):128-174, 2004.

[27] C. Powers, M. Schunter, "Enterprise Policy Authorization Language," Version 1.2, *W3C Member Submission*, Nov. 2003.

[28] L. Rosen, *Open Source Licensing: Software Freedom and Intellectual Property Law,* Prentice Hall 2005.

[29] G. Simon, "Accelerating Research through the National Health Information Network," Meeting Notes, FasterCures: the Center for Accelerating Medical Research, 7 Jan. 2005.

[30] Sweeney, Latanya. "k-anonymity: a model for protecting privacy." *Int'l J. of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5): 557-570, 2002.