Toward a Comprehensive Disclosure Control Framework for Shared Data

Scott E. Coull
RedJack, LLC.
Silver Spring, MD
scott.coull@redjack.com

Erin Kenneally
Elchemy
San Diego, CA
erin@elchemy.org

Abstract—**Although the goal of pervasive data sharing has persisted for over a decade, most large-scale efforts fail to reach the critical mass of participation necessary to sustain it due to the excessive costs involved. These costs often stem from the lack of standardized methodology and tools to implement disclosure controls that address data sharing risks. We present a framework that addresses the problem comprehensively by considering policy level risk (*e.g.,* NDAs) and technical (*e.g.,* data anonymization) disclosure control issues in concert. Doing so facilitates a superior balance of utility and risk mitigation by ensuring that policy and technical approaches complement one another. Moreover, the framework is driven by the pragmatic utility goals of the data release rather than general risk factors, which helps to focus the effort on exactly those parts of the data necessary to achieve desired goals. The output of the framework is a standardized audit trail and description of the data sharing scenario, which enables the reuse of key components in other data sharing efforts. The framework greatly decreases the data publisher's overall costs while simultaneously enabling a more evolved and effective balance between utility and risk management in data sharing.**

I. Introduction

Despite significant attention from policymakers and the research community, many initiatives to encourage pervasive data sharing have stalled due to security and privacy concerns. These data sharing efforts are hindered by a lack of standard methodology for setting utility goals, assessing risks, choosing appropriate disclosure controls, and implementing those controls to mitigate risks while maintaining utility. The current state of practice encourages the application of ad-hoc policy and technical approaches that often fail to provide an appropriate balance between the utility of the data and its risks. Moreover, the inherent variability in the disclosure control process among data sharing efforts makes it difficult to re-use legal or technological infrastructure, which results in excessive labor costs and an inability to properly audit the process. To achieve effective data sharing, a standardized process is necessary to guide the data publisher from setting utility goals all the way to implementing disclosure controls, while allowing the costs of this effort to be amortized across many data releases.

The disclosure control framework is made up of three basic components: templates, environments, and a risk assessment methodology. Generally speaking, a *template* is a data structure that encodes information about disclosure control components that transcend individual data releases or sharing scenarios. An *environment*, on the other hand, is a concrete instantiation of the disclosure controls and related data sharing infrastructure chosen by the publisher for a particular data release or scenario (*e.g.,* access controls, server software, etc.). Finally, the *risk assessment framework* guides the publisher through a decision-making process that essentially transforms a set of templates describing available controls and sharing options into fully-specified environments that can (given the right infrastructure) be automatically implemented and guaranteed to reflect the risk and utility goals of the release. A high-level overview of the workflow is shown in Figure 1. The publisher begins by establishing the primary utility goals of the release, which then inform the identification of associated risks to be mitigated. In the second phase of the workflow, the publisher chooses a set of disclosure controls to apply to the data, which may include controls to change the data or establish penalties for misusing it. The process ends by having the publisher describe how the utility goals and identified risks are impacted by the chosen controls. Throughout this process, the templates restrict the questions and options given to the publisher, and connect those options to the available environment configurations. Due to space restrictions, we provide only general specifications for the components and a high-level description of the assessment methodology itself, and refer interested readers to the full version of the paper for more details [1].

This approach offers a number of benefits over current ad-hoc methods. For one, we are able to reuse the basic components of past data releases through established templates; even going so far as to enable community-wide sharing of common templates. These templates describe the standard language of legal documents and basic functionality of technical disclosure controls methods, and then provide a common interface for customizing them. Additionally, the risk assessment methodology itself forces the publisher to consider utility and risk together while providing a unified set of both policy and technical options to achieve the goals of the data release, which enables a more refined balance between utility and risk mitigation. The outputs of each phase in the process also provide a standardized way of describing and justifying the data release so that it can be easily audited by third parties, like Institutional Review Boards (IRBs) or regulatory agencies. In some cases, this standardized information can be used to automatically configure data collection and sharing environments that can be verified to meet the goals set forth in the framework outputs. Overall, the framework provides a way to minimize the long-term costs of the disclosure control and data release process, while simultaneously providing data that is more useful and where the risks are more well-understood by all parties involved. In the remainder of this paper, we describe

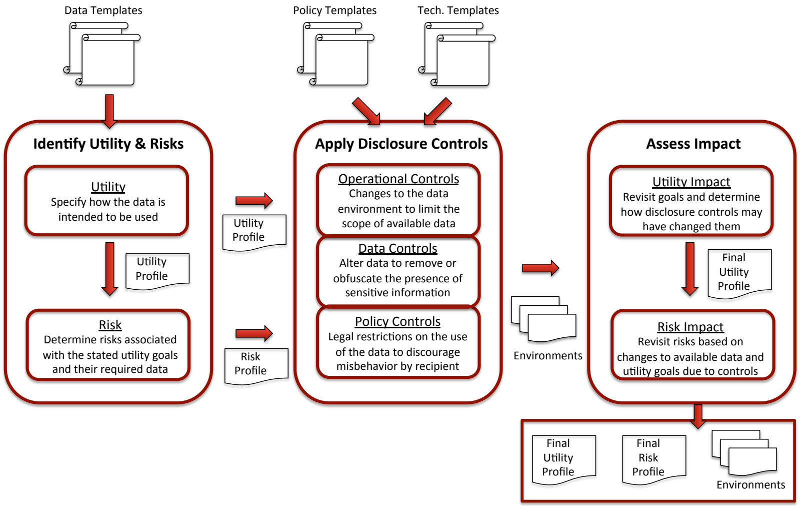

Data Templates
Policy Templates
Tech. Templates
Identify Utility & Risks
Utility
Specify how the data is intended to be used
Utility Profile
Risk
Determine risks associated with the stated utility goals and their required data
Utility Profile
Risk Profile
Apply Disclosure Controls
Operational Controls
Changes to the data environment to limit the scope of available data
Data Controls
Alter data to remove or obfuscate the presence of sensitive information
Policy Controls
Legal restrictions on the use of the data to discourage misbehavior by recipient
Environments
Assess Impact
Utility Impact
Revisit goals and determine how disclosure controls may have changed them
Final Utility Profile
Risk Impact
Revisit risks based on changes to available data and utility goals due to controls
Final Utility Profile
Final Risk Profile
Environments

Fig. 1. Overview of the disclosure control framework.

the components of the framework and their interactions with an eye toward computer network and security data, then conclude with a brief description of how the framework is being used in the FCC's on-going Measuring Mobile Broadband project [2].

II. Creating Reusable Templates & Environments

The foundation of the framework is built upon reusable components called templates and environments. The purpose of the templates is to clearly separate the baseline information and procedures that exists across all data releases from the information about a specific data sharing scenario that must be extracted from the publisher. Based on the specifics of the scenario at hand (*e.g.,* type of recipient, data, etc.), there may be many such templates to describe relevant legal and policy documents, as well as the format of the data being released and the applicable technical disclosure controls in use. All templates contain a distinguished name, a user-friendly description of the functionality of the template, and categorical information used to organize the templates for easy examination by the publisher. The key functionality here, is that the template must be able to facilitate the translation of the publisher's choices in the risk assessment and data sharing process into a specification for a concrete implementation. In this paper, we consider document, data, and technology templates, though other types may be added as the need arises. Examples for each of these templates, encoded in JavaScript Object Notation (JSON) format, can be found in Figure 2.

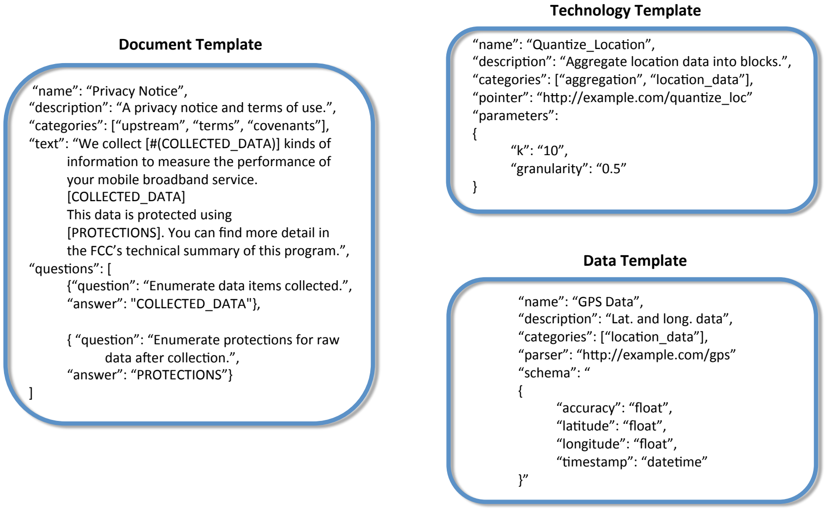
A *document template* is meant to encode the boilerplate text of legal and policy documents commonly used when collecting and sharing data, along with a series of questions that must be answered by the publisher to customize that text to the current scenario. The boilerplate text contains variables associated with each of the questions such that answering the questions fills in the blanks, so to speak, and allows us to create a complete document that can be used when collecting or sharing data. This is similar to the way privacy impact assessment templates are currently used, though on a much broader scale. The template also contains category information about the types of policy controls (*i.e.,* clauses) contained within the documents.

To describe the data being released, we use a *data template* that contains information on the syntax of the data type and how disclosure controls may be applied. More specifically, the data template consists of parsing rules (or a pointer to a parser implementation) and a data schema that breaks the data into individual fields that we may apply disclosure controls to. Each of these fields is associated with a type that is used to determine which disclosure controls may be used on that field.

The *technology templates* describe a single implementation of a disclosure control or supporting technology, such as server software or data collection utilities. In practice, these technology templates will often be abstractions of specific parts of much larger technologies or software implementations, such as a specific type of data filter in the collection software. The template includes information about the field and data types that it may be applied to, its disclosure control categories (discussed in Section IV), available parameters including default settings, and a pointer to its implementation. This information is enough to guide the application of controls to the appropriate types of data, and to the appropriate fields within that data.

The environments are simply sets of templates that have been chosen and configured by the publisher during the risk assessment and data sharing process. The primary purpose of these environments is to provide a concise and standardized description of the data sharing scenario, including specific instantiations of policy documents, data sharing software, and disclosure control parameter settings. Like the templates, we limit the scope of the environments we examine to only include collection and sharing environments, which are the two areas where disclosure controls are most often used for network and security data. Other situations where there is no control over the data being collected or which have intermediary processing steps may use a different number of environments.

The *collection environment* is made up of the completed templates that govern how the data is received from upstream providers, whether they be individual users or large organizations. This environment may include privacy policies, collection filters, and the storage format for the collected data. The *sharing environment* specifies the set of controls used



Document Template

```
"name": "Privacy Notice",
"description": "A privacy notice and terms of use.",
"categories": ["upstream", "terms", "covenants"],
"text": "We collect [#(COLLECTED_DATA)] kinds of
        information to measure the performance of
        your mobile broadband service.
        [COLLECTED_DATA]
        This data is protected using
        [PROTECTIONS]. You can find more detail in
        the FCC's technical summary of this program.",
"questions": [
        {"question": "Enumerate data items collected.",
        "answer": "COLLECTED_DATA"},

        { "question": "Enumerate protections for raw
                data after collection.",
        "answer": "PROTECTIONS"}
]
```

Technology Template

```
"name": "Quantize_Location",
"description": "Aggregate location data into blocks.",
"categories": ["aggregation", "location_data"],
"pointer": "http://example.com/quantize_loc"
"parameters":
{
        "k": "10",
        "granularity": "0.5"
}
```

Data Template

```
"name": "GPS Data",
"description": "Lat. and long. data",
"categories": ["location_data"],
"parser": "http://example.com/gps"
"schema": "
{
        "accuracy": "float",
        "latitude": "float",
        "longitude": "float",
        "timestamp": "datetime"
}"
```

Fig. 2. Example document, data, and policy templates.

when providing the data to a downstream recipient, such as a researcher, data repository, or the general public, and may include non-disclosure agreements (NDAs), authentication and access control mechanisms, and the disclosure controls to be applied to the data itself. With the right infrastructure, it may be possible to automatically generate the implementation of these scenarios using, for example, a set of baseline virtual machines. At the very least, the environments output by the framework provide a reasonable basis for auditing the publisher's data collection and sharing practices.

III. Utility & Risk Assessment

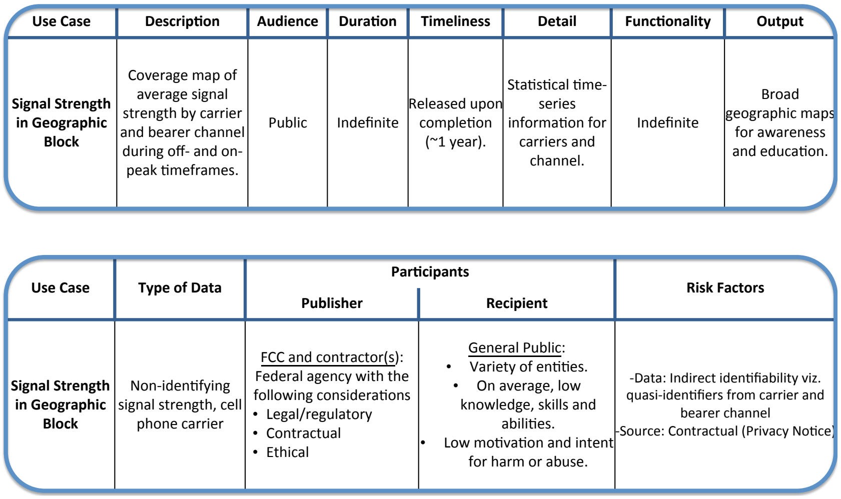
Utility and risk are often inextricably linked in data sharing efforts. Despite this close relationship, most risk assessment methodologies (*e.g.,* [3], [4], [5]) rarely, if ever, explicitly consider utility even though it is the driving force behind the effort. Rather than ignoring utility, our methodology makes it the central focus. The first step in our framework, therefore, is to assess the utility goals of the data release and the risks related to achieving those goals. These key factors are captured in *utility and risk profiles*, respectively. The profiles allow the data publisher to create a concise and standardized audit trail of the decision-making process underlying the data release. An example of these profiles is given in Figure 3.

Both the utility and risk profile reflect the *use cases* for the data. The *utility profiles* asks the publisher to provide a high-level description of the use case, as well as the requirements for standard properties of the data sharing scenario:

- Audience: The audience of the data release describes the type of community that is intended to act as the recipient of the data. The audience continuum can be roughly divided into individuals at the most restrictive end, consortiums in the middle, and public release at the most relaxed end.
- Duration: Once the data recipient is given access to the data, the duration criteria indicates how long the recipient may access that data. Data access durations may span from short-term access for real-time operational uses to indefinite access for general research purposes.
- Timeliness: Certain uses of the shared network data may place requirements on how quickly the data is made available to the recipient. Here, the continuum of setting choices ranges from real-time access to longitudinal data collection with long lag times between collection and availability.
- Detail: Another important aspect of utility is the level of detail required by the data recipient. Some operational tasks require detailed data about events or records, while general research use is most concerned about overall trends that manifest themselves in the data. Therefore, one may consider a continuum of data detail from event-specific data to general trend information derived from the data.
- Functionality: We may also consider the level of specificity associated with the data use cases. On one end of the functionality spectrum are data releases tailored to specific concrete tasks, while on the other are open-ended tasks.
- Output: The final aspect of data utility revolves around the intended outcome of using the data, or the output of the recipients' interaction with the data. Therefore, private knowledge lies on the most restrictive end of potential outputs, while publication would be the most broad form of dissemination.

These properties roughly cover the aspects of the data sharing scenario that the publisher can control and which affect the "success" of the data release effort with respect to how the data is ultimately used. During this part of the framework, the publisher also chooses one or more data templates that describe the type of data that is being collected, so that appropriate controls may be applied in the upcoming disclosure control selection phase.

While the utility profile helps helps determine what we



Use Case	Description	Audience	Duration	Timeliness	Detail	Functionality	Output
Signal Strength in Geographic Block	Coverage map of average signal strength by carrier and bearer channel during off- and on-peak timeframes.	Public	Indefinite	Released upon completion (~1 year).	Statistical time-series information for carriers and channel.	Indefinite	Broad geographic maps for awareness and education.

Use Case	Type of Data	Participants		Risk Factors
		Publisher	Recipient	
Signal Strength in Geographic Block	Non-identifying signal strength, cell phone carrier	FCC and contractor(s): Federal agency with the following considerations • Legal/regulatory • Contractual • Ethical	General Public: • Variety of entities. • On average, low knowledge, skills and abilities. • Low motivation and intent for harm or abuse.	-Data: Indirect identifiability viz. quasi-identifiers from carrier and bearer channel -Source: Contractual (Privacy Notice)

Fig. 3. Simplified version of utility and risk profiles.

need to protect, the *risk profile* specifies the risks identified for each of the use cases. These are the risks that ultimately need to be mitigated with disclosure controls while trying to ensure the utility requirements are met. The risk profile is made up of three properties that capture the relevant information necessary to explicitly describe the risks and their sources. These properties include:

- Type of Data: The type of data involved in the given use case may impose various obligations or restrictions related to the standard of care in collecting, using or disclosing it. For instance, Personally Identifiable Information (PII) has specific restrictions on its use imposed by laws and ethical risk sources. Mapping out the specifics types of data, from among all fields available in the relevant data templates, helps to identify the pertinent risk factors.
- Participants: The role of the characteristics of the publisher and recipient also alter the impact of the risk factors that may arise due to the data in the previous column of the profile. For the publisher, it is important to consider that there are some industries (*e.g.,* health care, finance, etc.) where there are specific regulations and professional standards that impose risk for sharing certain types of data. For the recipient, their overall potential for abusing the shared data should be considered, including motivating factors (*e.g.,* money, fame) and their technical expertise in bypassing any applied disclosure controls.
- Risk Factors: These are the actual identified risks that arise due to the combination of participants and data for the given use case. Each risk factor is derived from risk sources that include laws, private agreements, proprietary rights, ethical obligations, unilateral policies (*e.g.,* Terms of Use), and best practices. The combination of data and participants should be evaluated against each of the categories for the data sharing scenario in question, and relevant risk factor should be listed with their source.

We note that it is difficult, or perhaps impossible, to provide a meaningful quantitative risk assessment for general data sharing settings. In particular, there is no clear definition of utility to allow for any integration into such a framework, and beyond that we must consider that the policy landscape constantly changes, that many quantitative measures do not capture the true risk for some types of data [6], [7], and that risk is often a subjective and relative notion based on the context. The overarching philosophy behind our approach is based on the idea that explicitly stating the factors considered during data release allows for transparency that is itself a mitigating factor when considering the inevitability of changes in the future. Thus, the goal is not to guarantee any particular level of safety, but to provide the information needed for the publisher and other participants to make educated choices about data sharing.

IV. Choosing Disclosure Controls

Once the utility and risk profiles have been created in the assessment phase, the data publisher's next task is to create one or more environments to sufficiently mitigate the identified risks and uphold the utility goals. This is accomplished by choosing from among the available document and technology templates, then configuring their properties and applying them to the data as a whole, or in part. In many cases, there are several ways to reach the same end state by applying different combinations of controls, such as when we limit data collection via filters or delete parts of the data after it is collected. Obviously, the chosen controls may achieve the same level of risk mitigation, but often at different costs. The category associated with each template is used to organize the available options, while the name and description provide an understanding of the specific usage, benefits, and drawbacks for the template. There are three broad classes of that span operational, data, and policy controls, along with several subcategories for each.

Operational controls restrict different aspects of the supporting data collection and sharing infrastructure in an effort to minimize the exposure of the sensitive data, either before it is collected or after it is made available to the recipients. Some examples include access controls, use of specific data formats, and timing restrictions on data availability. The operational controls are broken into the following six sub-categories:

- Filtering: Limits the data to a specific sub-population as it is being collected and stored. In some cases, the most risky data population can simply be ignored during collection to mitigate its potential risks.
- Duration: Specifies the amount of time the data is available to recipients, which limits exposure of the data to potential abuse.
- Timeliness: Controls how long the data is retained before it is made available to the recipient. Based on the time-sensitivity of the risk factors involved in the use cases, it may be possible to enforce long waiting periods before the data can be accessed.
- Length: Collecting data over short periods of time often provides more limited exposure for potentially risky data, while longitudinal collection often leads to information that is deeply rooted in the patterns that emerge over time, which can make mitigation more difficult.
- Format: Some data formats naturally encode less detailed information than others, or may naturally restrict the data to only a small number of pertinent fields. This may help to focus the data collection effort to only the most basic information necessary.
- Access: There are several methods that can be used to control and audit access to the data itself, including limited query interfaces or other mitigated environments. When the data is accessed within a controlled environment, it may be possible to offset any potential risk factors of data exposure with stronger authentication and auditing mechanisms to recover from malicious activities.

Data controls alter the data itself after it has already been collected and stored. Here, the data controls may be applied to the dataset as a whole, to specific rows or columns, or to very specific pieces of information (*i.e.,* data cells). These data abstractions are dictated by the data templates associated with this data release. There are six sub-categories of data controls:

- Deletion: Simply removes a row (record), column (field), or specific set of cells in the data. The difference between deletion and the operational limitations above is that using deletion allows the publisher to examine the data and make more dynamic disclosure control choices based on the results of the collection itself.
- Aggregation: Takes the values of a field over several records and aggregates them into a single record value. For instance, the age of participants in a survey may be aggregated by taking the average of their ages. These methods attempt to blend the impact of any one record in with others while still providing useful trending information.
- Generalization: Uses the semantics of a field to generalize several related classes of values into a single large class. An example of this would be truncating zip codes or Social Security numbers, which effectively generalizes the values into groups based on larger geographic areas.
- Pseudonymization: Replaces identifiers with a linkable or partially-unlinkable pseudonym to hide the real identity associated with the record, but maintain the ability to group those records together.
- Perturbation: Changes the value of a field and combining it with noise, such as adding noise taken randomly from a Laplace distribution to an number.
- Synthetic Data: Given a set of specific statistical properties to maintain, generative models can be trained to produce data that is guaranteed to match those properties, but which has no connection to the original data for any other property. Data imputation techniques are also considered to be a type of synthetic data generation method.

Policy controls mitigate risk not by trying to hide or limit access to the risky data, but instead by providing strong incentives for appropriate behavior and penalties for abuse. In addition to risks arising from the exposure of the data itself, there are often other risks related to various policy aspects of the collection and sharing process, such as the need for informed consent from users or transitive application of agreements from upstream providers to downstream recipients. Examples of these policy controls include privacy policies, memorandums of agreement, data licenses, and non-disclosure agreements associated with upstream data providers. The basic functionalities of the policy controls are categorized as follows:

- Performance: Describes the bargain or exchange, such as the scope of data that is protected, and license grants or restrictions.
- Consideration: Restrictions related to required fees or necessary services that are related to the collected data or its source.
- Covenants & Conditions: Requirements or obligations placed on the parties for use of the data, such as consent notices, confidentiality obligations, and destruction of data after a specified period of time.
- Accountability & Enforcement: Guarantees and mechanisms to enforce or police them, including penalties and auditing rights.
- Terms & Termination: Specific termination conditions for the use of the data, time period of use, conditions under which can it be ended.
- General: Basic requirements imposed by governing law or other third-party governing law, interpretation and adjudication; binding effects, third party beneficiaries.

The process of choosing the disclosure controls using this framework is guided by the categorization of the policy and technology templates, and their applicability to the chosen data templates. Once a control is selected by the publisher, the list of questions or parameters found in the associated template are presented so that they may be customized to the risk level of the current data sharing scenario. The publisher also chooses which environment (if more than one exists) the control should be associated to. The output of this phase of the framework is a set of environments containing the configured templates chosen by the data publisher. While the framework does not currently consider the notion of completeness (*i.e.,* the idea that there is a necessary set of templates), it is possible that in the future it may be extended to establish requirements for certain types of controls. For instance, sharing obviously cannot occur without choosing some type of server technology, and so that may end up becoming a requirement in future iterations of the framework's implementation.

V. Evaluating Disclosure Control Impact & Using Framework Outputs

The final phase of the framework comes full circle to determine how the choices of disclosure controls has changed the original utility goals and risk factors identified at the start of the process. The evaluation proceeds by adding an additional column to the utility and risk profiles, called the *impact statement*. The publisher uses the field to specify how they believe things have changed, either quantitatively or qualitatively. As with the initial profiles, we cannot rely on a one-size-fits-all approach when talking about quantitatively measuring change in inherently qualitative utility properties and risk factors. We can, however, make some quantitative measurements where they naturally occur, such as an increase in lag time between the time data is collected and when it is made available to recipients. In general, though, we believe that qualitative impact statements will be the most generally useful approach. Again, there is no claim that the data is guaranteed to be safe, but these final profiles help encourage defensible and pragmatic solutions. In fact, use of the development of community-driven templates and use of the framework itself helps to set a standard for what can be considered to be a "reasonable" level of due diligence on the part of the publisher.

Once the profiles have been completed, the framework outputs the final utility and risk profiles, as well as the set of environments created during the disclosure control phase of the framework. There are several uses for these outputs that greatly improve the current state-of-practice. Probably the most obvious use is to provide the profiles and environments to third-party auditors to review the decisions made in choosing the controls. Since the profiles provide direct support for the environment configurations and those environments are standardized, it is much easier to have a data privacy expert or attorney verify that the controls meet the necessary requirements. By comparison, the current approach would be to engage the experts on an ad-hoc basis with little or no information about the complete data sharing scenario, instead receiving only piecemeal verification of the controls and data sharing policies.

Another, more ambitious use of the output is to use it to automatically assemble implementation artifacts for each of the environments. As mentioned earlier, it is possible to create one baseline virtual machines (VMs) for each environment in the output, with all of the available tools pre-installed. Then, the applications within the VMs are configured according to the templates within their respective environments. Such a system would remove almost all technology and implementation costs involved with new data sharing efforts, and enable simple verification procedures for compliance with stated policies. At the moment, compliance checking of any sort is actually impossible because the disclosure control policies are not formalized and the implementations are not standardized.

VI. Conclusion

The disclosure control framework presented in this paper has been used, in part, to develop privacy and disclosure policy recommendations for the FCC's Measuring Mobile Broadband (MMB) project [2]. The main goal of the project is to gather information about the speed, performance, and coverage of current mobile carriers within the United States from the mobile devices of volunteers. To accomplish this goal and make the results available to the public, it is important to protect sensitive information about the volunteer's location information while still ensuring that the level of specificity for each measurement provides useful data for distinguishing performance characteristics among different geographic areas. The situation is complicated by the sheer number of entities involved with the data sharing effort – mobile carriers, volunteers, the FCC, independent researchers, and curious members of the general public. Throughout the on-going process, we have used the framework, as illustrated in the examples throughout this paper, to clearly and coherently organize the information necessary to make appropriate decisions about disclosure controls. Without such a framework, the sheer complexity of the situation would undoubtedly allow for gaps in the disjointed efforts by engineers, the legal community, and privacy experts in addressing the utility, policy, and privacy concerns arising from this data release.

Acknowledgments

This work was supported as part of the PREDICT project operated by the U.S. Department of Homeland Security's Science & Technology Directorate.

References

[1] S. E. Coull and E. E. Kenneally, "A Qualitative Risk Assessment Framework for Sharing Computer Network Data," 2012 TPRC, Tech. Rep., March 2012, available at http://dx.doi.org/10.2139/ssrn.2032315.

[2] Federal Communication Commission, "Measuring Mobile Broadband," Accessed at: http://www.fcc.gov/encyclopedia/measuring-mobile-broadband.

[3] J. Domingo-Ferrer and V. Torra, "Disclosure risk assessment in statistical microdata protection via advanced record linkage," *Statistics and Computing*, vol. 13, no. 4, pp. 343–354, 2003.

[4] W. E. Yancey, W. E. Winkler, and R. H. Creecy, "Disclosure Risk Assessment in Perturbative Microdata Protection," in *Inference Control in Statistical Databases*, ser. Lecture Notes in Computer Science, J. Domingo-Ferrer, Ed., 2002, vol. 2316, pp. 135–152.

[5] K. E. Emam and F. K. Dankar, "Protecting Privacy Using k-Anonymity ," *Journal of the American Medical Informatics Association* , vol. 15, no. 5, pp. 627 – 637, 2008.

[6] S. Coull, F. Monrose, M. Reiter, and M. Bailey, "The Challenges of Effectively Anonymizing Network Data," in *Proceedings of the DHS Cybersecurity Applications and Technology Conference for Homeland Security (CATCH)*, March 2009, pp. 230–236.

[7] M. A. Rothstein, "Is Deidentification Sufficient to Protect Health Privacy in Research?" *The American Journal of Bioethics*, vol. 10, no. 9, pp. 3–11, 2010.