Abstract—Widespread sharing of scientific datasets holds great promise for new scientific discoveries and great risks for personal privacy. Dataset handling policies play the critical role of balancing privacy risks and scientific value. We propose an extensible, formal, theoretical model for dataset handling policies. We define binary operators for policy composition and for comparing policy strictness, such that propositions like “this policy is stricter than that policy” can be formally phrased. Using this model, the policies are described in a machine-executable and human-readable way. We further present the Tags programming language and toolset, created especially for working with the proposed model. Tags allows composing interactive, friendly questionnaires which, when given a dataset, can suggest a data handling policy that follows legal and technical guidelines. Currently, creating such a policy is a manual process requiring access to legal and technical experts, which are not always available. We present some of Tags’ tools, such as interview systems, visualizers, development environment, and questionnaire inspectors. Finally, we discuss methodologies for questionnaire development. Data for this paper include a questionnaire for suggesting a HIPAA compliant data handling policy, and formal description of the set of data tags proposed by the authors in a recent paper.

I. INTRODUCTION

Wide dissemination of datasets holds great promises for science — findings can be corroborated, research cost reduced through data reuse, and new studies are made possible by combining existing data into new datasets, to name a few. But wide dissemination of datasets also poses risks to the data subjects. Privacy of human subject has to be respected. Precise locations of endangered species populations, rare minerals, or ancient ruins should not be easily available to poachers, illegal miners and tomb raiders. Many laws, regulations and best-practices were formed in order to balance the positive and negative potential of data sharing. In the US alone, there are more than two thousand rules and regulations governing data sharing [15]. To share data while respecting legislation, the scientific community has resorted to specialized data repositories, e.g., repositories specifically designed for medical data. While legally sound, this approach leads to fractured data storage infrastructure, and leaves datasets that contain data governed by certain law combinations without a repository.

In [16], the authors propose the concept of datatags as a way of ensuring that the handling of, and the access requirements to a dataset are commensurate with the risks of harm it poses. When deposited in a datatags-compliant data repository, a dataset is associated with a datatag which defines a machine-actionable policy under which the dataset should be handled. This ensures that the data repository handles the dataset properly, from a legal and contractual standpoint. By limiting the amount of datatags to a few well-defined choices, managing, reasoning about, and implementing robust datatags-compliant repositories becomes easier. Finally, a sample set of datatags is suggested (see Figure 1).

Datatags-compliant repositories will help automate open science, and facilitate data sharing [13]. Furthermore, since Datatags-based systems can predict the legal and technological requirements for the collection and handling of a dataset, they can also serve other use-cases, such as IRBs and research design.

Once a set of datatags has been defined, two challenges need to be addressed. First, a tag has to be formally described, in a way that allows formal reasoning about the data handling policies the tags imply, and clear transition into implementation. Second, a mechanism to help users match a dataset with the datatag most appropriate for it has to be created. Such matching requires familiarity with the dataset’s history and collection methods, and legal and technological expertise — the latter two are not always readily available to researchers. Enforcing the policy detailed by the datatag is a challenge in its own right, is not in the scope of this paper.

In order to be useful, a datatag matching system should have

<table>
<thead>
<tr>
<th>Tag Type</th>
<th>Description</th>
<th>Security Features</th>
<th>Access Credentials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>Public</td>
<td>Clear storage, Clear transmit</td>
<td>Open</td>
</tr>
<tr>
<td>Green</td>
<td>Controlled public</td>
<td>Clear storage, Clear transmit</td>
<td>Email- or OAuth Verified Registration</td>
</tr>
<tr>
<td>Yellow</td>
<td>Accountable</td>
<td>Clear storage, Encrypted transmit</td>
<td>Password, Registered, Approval, Click-through DUA</td>
</tr>
<tr>
<td>Orange</td>
<td>More accountable</td>
<td>Encrypted storage, Encrypted transmit</td>
<td>Password, Registered, Approval, Signed DUA</td>
</tr>
<tr>
<td>Red</td>
<td>Fully accountable</td>
<td>Encrypted storage, Encrypted transmit</td>
<td>Two-factor authentication, Approval, Signed DUA</td>
</tr>
<tr>
<td>Crimson</td>
<td>Maximally restricted</td>
<td>Multi-encrypted storage, Encrypted transmit</td>
<td>Two-factor authentication, Approval, Signed DUA</td>
</tr>
</tbody>
</table>

Fig. 1. Blue to Crimson model set of datatags, proposed in [16]. The Blue tag is appropriate for datasets that pose no risks. As the risk level increases, so does the required access credentials, the security imposed by the handling requirements, and the strictness of the DUA terms and execution.
the following traits:

- **Algorithmic** Require input from the data depositor only.
- **Deterministic** Given the same input the system should yield the same results (assuming no legislation or technological changes).
- **User friendly** Deciding on the datatag of a dataset, and hence on the data handling policy applied to handle it, has ethical and legal implications. Thus, data depositors need to be well informed during the process, and user errors should be minimized.
- **Maintainable** The system captures non-trivial legal and technological knowledge, that has to be composed and maintained by human domain experts.
- **Assume reasonable user knowledge** Data depositors should not be expected to have legal or technological expertise. They are expected to be familiar with the way the dataset in question was collected.

This paper will propose an answer to both challenges, defining data handling policies as points in a multi-dimensional space, and creating a datatag decision aid using a decision graph. The rest of this paper is organized as follows: Section II proposes a formal description of data handling policies, and uses this description to define datatags. Section III describes Tags’ tools and briefly discusses methodologies for using it. Sections VI and VII discuss related and future work.

II. POLICIES AS SPACES

A given a data handling policy can be described as composed from many independent aspects. For example, the requirement to encrypt data ‘at rest’ is independent of the requirement to have data users sign a data use agreement (DUA), or even from the requirement to encrypt the data on transmission. Granted, a policy that requires a signed DUA and an encrypted storage is likely to also require encrypted transmission. It is possible, however, to create a policy that requires encrypted storage and signed DUA, but allows clear transmission. It is therefore useful to allow description of such policies, if only to create an algorithm that warn against them.

Each of the data handling policy aspects has multiple possible requirements. For example, a policy might allow accepting a data use agreement in an implied manner, require using a click-through agreement process, or require a signed document. It might allow storage to be in the clear, encrypted, or even multi-encrypted. These possible requirements can be ordered from lenient to strict. In cases where the possible requirements of a data handling policy aspect cannot be ordered, it is possible to either generalize them, or split the aspect. For example, verifying user registration via an email address or an OAuth account is equivalent, as both demonstrate an account on an external system. Thus, as possible requirements for VerifyRegistration, email is no stricter than OAuth, nor is OAuth stricter than email. The solution is to replace them with a more general usingExternalSystem possible requirement. The actual choice between email, OAuth, or other equivalent technologies is then left to data repository system implementors.

An example for an overly broad aspect definition is Encryption. As there are two possible aspects to encrypt — storage and transmission — the possible requirements transferOnly and storageOnly have no strict ordering. Splitting the Encryption aspect to StorageEncryption and TransmissionEncryption solves this issue.

We propose viewing the independent aspects of data handling policies as orthogonal axes of an ordinal space, where the possible requirements of a given aspect are the coordinates along its axes. We call this space “Data Handling Policy Space” (or DHP space for short). Viewed like this, data handling policies become points in a DHP space. See Fig. 2 for a 2D example.

More formally, given independent aspects $a_1, a_2, \ldots, a_n$, where each aspect $a_i$ has $k_i$ possible requirements $r_1^i, r_2^i, \ldots, r_{k_i}^i$, a data handling policy $P$ is a point in an $n$-dimensional data handling policy space, and is specified by the n-tuple:

$$P = \langle r_1^1, r_2^2, \ldots, r_n^k \rangle$$

Where $1 \leq l_i \leq k_i$. Aspect order is arbitrary, but of course has to be consistent once determined.

A. Relations Over Policies

Building on the above description of a data handling policy, we now define equality and partial order, based on strictness.

**Equality** Given policies $P = \langle p_1, \ldots, p_n \rangle$ and $Q = \langle q_1, \ldots, q_n \rangle$, defined in the same DHP space, we say that $P$ is equal to $Q$ iff all their requirements are equal. Formally:

$$P = Q \iff \forall i \in [1..n], p_i = q_i$$

**Strictness** Given DHP space $S$ defined over $n$ aspects, and policies $P = \langle p_1, \ldots, p_n \rangle$ and $Q = \langle q_1, \ldots, q_n \rangle$ defined in $S$, we say that $P$ is stricter than $Q$ iff along all aspects of $S$, $P$ has requirements that are equal or stricter than $Q$’s, and on at least one aspect that value is strictly stricter. Formally:

$$P > Q \iff \forall i \in [1..n], p_i \geq q_i \land \exists i \in [1..n], p_i > q_i$$

For any DHP space with more than a single aspect, strictness is a partial order relation.

**Derived** Using equality and strictness, the relations more lenient ($<$), stricter than or equal to
more lenient than or equal to ($\leq$) are intuitively defined.

**Composition**

Given DHP space $S$ defined over $n$ aspects, and policies $P = \langle p_1, \ldots, p_n \rangle$ and $Q = \langle q_1, \ldots, q_n \rangle$ defined in $S$, $P$ and $Q$ can be composed to a new policy, by selecting the stricter requirement for each aspect. Formally:

$$P \oplus Q \triangleq (\max(p_1, q_1), \ldots, \max(p_n, q_n))$$

Where $\max(f_1[a], f_1[b])$ is defined as $f_1[a]$ if $f_1[a] > f_1[b]$, and as $f_1[b]$ otherwise. Note that by definition $P \oplus Q \geq Q$ and $P \oplus Q \geq P$.

**B. Compliance and Support Sub Spaces**

Security and access restrictions are not breached by over-doing. For example, if a dataset is allowed to be transmitted in the clear, it is also acceptable to encrypt it during transmissions. Formally, if a dataset is required to use policy $P$, a dataset repository may handle it according to any policy $Q$, when $P \leq Q$. Thus, each policy $P$ in DHP space $S$ defines two sub-spaces, demonstrated in Figure 2:

- **Compliance Space** All policies that do not breach $P$:
  $$\text{compliance}(P) \triangleq \{ Q \in S \mid P \leq Q \}$$

- **Support Space** All policies that $P$ does not breach:
  $$\text{support}(P) \triangleq \{ Q \in S \mid Q \leq P \}$$

A dataset requiring policy $P_1$ can be handled using any policy in $\text{compliance}(P_1)$, without breaching any laws. A data repository implementing policy $P_2$ can properly handle any dataset which requires a policy is in $\text{support}(P_2)$. In order to decide whether dataset $D$, requiring policy $P_D$, can be stored in repository $R$ under policy $P_R$, it is enough to require that $\text{compliance}(P_D) \cap \text{support}(P_R) \neq \emptyset$. When this assertion holds, we can say that $R$ can handle $D$ using policy $P$, writing $R \vDash_P D$. In the interest of open science, it is better to use the most lenient policy within that intersection.

Statistical disclosure limitation tools and sharing a differentially private version of the dataset [8], [9] can be viewed as moving the dataset to a more lenient point in DHP, at the expense of some of its aspects, such as its level of detail.

**C. DataTags and DHP**

Under the Data Handling Policy Space model, DataTags are a set of points in a defined DHP space, totally ordered by strictness. The compliance and support spaces of the tags are hierarchically contained. For the set of tags defined in [16], $\text{compliance(blue)}$ is the outermost compliance space, which serves the intuition that handling an open access dataset as if it required all its users to be fully accountable (red tag) will not create a security breach (it would, however, unnecessarily complicate the dataset handling). The outermost support space is $\text{support(crimson)}$, serving the complementary intuition.

The code addendum for this paper [5], contains the datatag set proposed in [16] phrased using this approach.

**III. TAGS, A GRAPH-BASED DATATAGGING SYSTEM**

We now present Tags, domain-specific programming language for defining data handling policy spaces, and creating friendly interactive questionnaires, that helps users arrive at a proper policy for a given dataset. We will also look at some of the tools available for Tags programs. Tags is an open source project under active development, so this paper covers only its basic concepts. For an up to date information, we refer the reader to the project website [4].

The DataTags concept serves multiple use cases, such as individual researchers, research groups, IRBs, and scientific and commercial data repositories [16]. Thus, the Tags parser and runtime engine are available in an embeddable Java library. This Library is currently embedded in two systems: CliRunner, a Java console application which serves as a development aid, and TaggingServer, a web application for conducting interviews, written in Scala.

A Tags program is called a questionnaire. An interview is the act of a user interactively answering questions from a questionnaire. Questionnaires are modeled after an interview with an expert, where the expert asks questions using a language the data depositor can understand, while writing comments in her notebook using professional, accurate, language. A Tags questionnaire is composed of two components: a tag space defining the professional language, and a decision graph, defining the questions and the flow of the interview. Tag space is similar to data handling policy space, but contains additional axes for auxiliary assertions, such as regulatory information and internal program variables. These additional assertions can be used later, e.g. for automatic customization of data use agreements. Different decision graphs may refer to the same tag space. Consequently, tag spaces may be used to define
standards that data repositories can comply with.

Tags was designed to be used by domain experts rather than programers. While it does require its users to adhere to formal syntax, it is a high-level language, and does not require handling memory allocation or read/write operations. To better serve the questionnaire development process, Tags supports top-down design, by allowing parts of the questionnaire to be marked as TODO.

A. Tag Space

A direct description of multi-dimensional space — namely, a list of dimensions and the possible values along them — may be hard to work with, especially as the number of described aspects increases. Furthermore, such list does not allow for logical grouping, and hinders collaboration, both are crucial for engineering a questionnaire. Thus, Tags describes tag spaces using named slots, and defines a simple algorithm for generating a list of dimensions from a set of slots. Tags supports four types of slots:

1) Atomic Slot: A slot containing a single atomic value. Values defined for this type of slot are totally ordered. Slots of this type are defined using the one of keyword.

   Storage: one of clear, encrypt, multiEncrypt.

2) Aggregate Slot: A slot that contains a set of values of its item type. Aggregate slots can be used to logically group multiple binary dimensions. They are defined using the some of keyword.

   ProtectedDataSubjects: some of livingPersons, deceasedPersons, endangeredSpecies, rareMinerals.

3) Compound Slot: Slots of this type consist of other slots, referred to as sub-slots. Compound slots do not define tag space dimensions directly. Rather, they are used to logically group other slots, and are defined using the consists of keyword.

   Handling: consists of Storage, Transit, Authentication.

4) Todo Slot: A placeholder slot, containing no values. Slots of this type are useful for building tags spaces using a top-down approach. In the tag space below, the intellectual property aspects of a the policy will be defined later:


   To improve user friendliness, slots and values can have an associated explanatory text. Unlike code comments, these texts are retained at runtime, and interview softwares can make them available to the interviewee. Explanatory text association is done by adding it, between square brackets, immediately after the name of the defined entity, like so:

   Storage [The way data are stored on the server]: one of clear, encrypt [Encryption by the server, before storage], multiEncrypt [Encryption by both the client and the server].

Fig. 3. The tag space defined in Listing 2, visualized by Tags’ CliRunner tool. Atomic slots are visualized as scales, with the most lenient value on the left. Aggregate slots are shown as baskets, containing some of the aggregate values, visualized as eggs. Compound slots are visualized as circles with outgoing arrows. The arrows are labeled with the name of the slot they point to.

Slots are statically associated with the type implied by the slot definition. For example, the type of the Storage slot defined above contains the values clear, encrypt and multiEncrypt. It is an error to assign a value to a slot, if said value is not a member of the slot’s type.

The root slot of the tag space definition is a compound slot called DataTags. Unfolding a slot-based tag space definition to a list of dimensions is done as follows:

1) Start with a list containing the DataTags slot.
2) While there are compound slots in the list:
   a) Remove the first compound slot, c
   b) For each of c’s sub-slots (except the TODO ones), prefix the name the sub-slot with c’s name and a delimiter\(^3\), and add the sub-slot to the list.
3) For each aggregate slot a in the list:
   a) Remove a from the list
   b) For each of a’s values, generate an atomic slot with the value’s name prefixed by a’s name, and a delimiter. The generated slot has two values, the lesser representing the value not present in the aggregate slot, and the greater representing the value’s presence. This pair of values can be thought of as YES and NO, or applies and does-not-apply, depending on the semantics of the values in aggregate slot.

When the algorithm terminates, the list of slots will contain only atomic slots with unique names, listing all dimensions of the described tag space, and the possible values for each dimension. The code addendum [5] offers a complete example for unfolding a tag space definition.

B. Decision Graph

The “program” part of a Tags questionnaire is called a decision graph, and is modeled after a graph with various type of nodes. These types denote program instructions for asking questions, setting tag values, and control flow. During execution — interview, in Tags terms — the runtime engine traverses the decision graph, executing the command in each

\(^3\)A delimiter is a character not allowed in a slot’s name.
node. The runtime maintains two data structures: Tags value, a single instance of type DataTags, as defined in the tag space part of the questionnaire, and a call stack. Thus, a Tags interview can be viewed as an interactive, stacked traversal of the decision graph. At the end of the interview, the engine’s Tags value contains the resultant data handling policy, and auxiliary assertions, if any.

Nodes are written between square brackets, and consist of a head and a body, separated by a colon (see Listing 1). The node head contains the node’s instruction (e.g. ask, call) and a possible id. The node body, if present, contains additional information required for executing the node’s instruction. For example, the body of a call contains the id of the node to call.

Listing 1. Sample nodes, showing various nodes structures: with and without id (lines 1 and 2, respectively), and a node with no body (line 3).

```
[>anId< set: storage=encrypted]
[end]
```

Tags support various node types, some of which are listed below.

a) **Set Node:** Assigns values to some of Tags Value’s slots. Atomic slots are assigned to using =, while aggregate slots are added to using ⊕. More than a single value can be assigned at a time, by separating the assignments using semicolon. For brevity, unique suffixes of a slot name can be used, instead of the full path to it. This type of abbreviation is demonstrated by the below example in the assignment to the HIPAA aggregate slot.

```
[set: DataTags/Handling/Storage=clear; HIPAA+=[expertDetermination, waiver]]
```

A set node can only make values stricter — in effect, it has the semantics of the ⊕ operator, defined in II-A. This prevents different sections of an interview from lifting each other’s restrictions: Consider a dataset containing mundane health data and private educational data. The educational data section of the questionnaire will conclude that the dataset has to be encrypted at rest. However, as the dataset does not contain private health data, the health section of the same questionnaire might allow it to be stored in the clear. Clearly, the dataset has to be stored encrypted. By preventing set nodes from replacing strict values with lenient ones, the health section cannot interfere with the decision made by the educational records section.

b) **Ask Node:** Present a question to the user, and wait for an answer. An ask nodes define a set of possible answers. ask nodes have a recursive structure: each answer has its own decision sub-graph, executed by the engine if that answer is selected.

```
[ask:
  {text: Do the data contain personally identifying information of humans?}
  {terms:
    {personally identifying information: This means the name, address, fingerprints...}}
  {answers:
    {yes, living: [set: livingPersons=yes]}
```
Listing 3. Sample decision graph for handling educational records.

```plaintext
1 [ask: Do the data concern humans?]
2   {yes: [set: Assertions=humanData]}
3       [ask: Does the data contain educational records?]
4         {answers:}
5             {yes: [call: eduCompliance]}
6             {no: [set: Transit=encrypt; Storage=encrypt]}<<-- defaults
7         [todo: Handle IP issues here]
8       [end]
9 !! Educational Compliance Section >>
10   [eduCompliance<< ask: Was written consent obtained?]
11   {answers:}
12       {yes: [reject: Cannot handle educational records without written consent.]}
13       {no: [set: Assertions=educationalRecords][1]}]
14 [end]
```

Listing 2. Tag space for the sample questionnaire.

DataTags: consists of Handling, Assertions, IP.
Handling: consists of Storage, Transit.
Assertions: some of humanData, educationalRecords.
Storage: one of clear, encrypt, multiEncrypt.
Transit: one of clear, encrypt.
IP[Intellectual Property]: TODO.

The interview decision graph is shown in Listing 3 and visualized in Figure 4. It has two components, the main one (lines 1–10) asks the user about the data subjects and, in case they are human, about the data type. Line 11 sets a default data handling policy — note that since set nodes cannot replace strict values with lenient ones, if the Storage slot contains encrypt or multiEncrypt, the set node at line 11 has no effect. Line 11 ends with an end-of-line comment: all text starting with <-- is ignored up to the end of the line.

The second component of the decision graph, starting at line 15, ensures the dataset was collected after the data subjects have signed a consent form. If that was not the case, the dataset is rejected (lines 18–19; again, the actual law is more complicated than this). Line 14, above the consent section, has a block comment documenting that section. Block comments, surrounded by <-- , can span multiple lines.

V. TOOLS AND METHODOLOGIES

Tags is a new language, but already has some tools for creating questionnaires and performing interviews. This section will introduce some of these tools, and discuss budding methodologies for questionnaire development.

A. Tags CliRunner

The main tool for developing questionnaires is CliRunner. This is a "swiss-army knife" command-line application. First and foremost, it allows running interviews. For debugging a questionnaire, CliRunner supports call stack and interview trace inspections. Code validations for finding unreachable nodes, unused tag values and invalid call nodes are also available. With the aid of Graphviz [10], CliRunner can create visualizations of tag spaces and decision graphs. Figures 3 and 4 were generated this way. CliRunner is built using an open architecture to support easy extension. Like the rest of the Tags code, it is open source.

B. Decision Graph Queries

Given a decision graph, it is possible to find all runs that yield policies with specific traits. For example, one can ask "which sets of answers result in a policy that allows clear transmission, and an assertion that the dataset holds sensitive data?". This type of question is useful, since the answer for most valid questionnaires should be "no such set of answers exists". CliRunner’s find-runs command allows running such queries, by finding all the interviews (i.e. sequences of nodes and answers) that result in a tag value that is a superset of the supplied tag (tag value a is a superset of tag value b if they both agree on the values in b’s non-empty slots). Listing 4 shows the find-runs output for a query over the sample questionnaire presented in Section IV. We hope to expand this type of queries to a richer set of tools that will allow formal questionnaire verification.

Listing 4. CliRunner's find-runs command can find all the runs that result in a superset of a given result.

```
answer (? for help): \find-runs Storage=clear
Run 1:
>[#1]< ask: Do the data concern humans?
-yes->
>[#2]< set
>[#3]< ask: Does the data contain educational records?
-no->
>[#5]< set
>[#6]< set
>[#7]< todo: Handle IP issues here
>[#8]< end
Final Tags:
DataTags/Assertions = humanData
DataTags/Handling/Storage = clear
DataTags/Handling/Transit = encrypt
```

Fig. 4. The decision graph in Listing 3, visualized by Tags’ CliRunner tool.
C. Web-based Interviews Using TaggingServer

TaggingServer is a web-based questionnaire runner, allowing users to go through a user-friendly interview online. Providing a good user experience for an interview is important, as answering a questionnaire which has legal and ethical implications might be a daunting experience.

TaggingServer provides application program interface (API), allowing it to be integrated into dataset depositing workflows of online data repositories, such as Dataverse [12] [7] [6]. The integration works as follows:

1) During the depositing process, the data repository requests an interview from a TaggingServer. The TaggingServer replies with a URL for the interview.
2) The repository redirects the user to the URL sent from TaggingServer.
3) The user is interviewed by the TaggingServer.
4) After the interview, the TaggingServer sends the result to the data repository. The data repository replies with a URL to redirect the user to.
5) The TaggingServer redirects the user to the provided URL. The data depositing process continues.

One important aspect of this integration is that the interview result is sent to the data repository directly from the TaggingServer, rather than manually by the user. This prevents users from fiddling with the interview result and from depositing one dataset under with the interview result of another. After the TaggingServer supplies the datatag to the repository, it is up to the repository to ensure that the dataset is handled according to the supplied tag, e.g. by using a sticky policy approach.

A TaggingServer hosting a questionnaire for creating data handling policies compliant with educational, health, and government records regulations, is available from http://datatags.org. Note that the tag space and questions were created as a proof of concept, demonstrating Tags’ ability to handle large questionnaires that span multiple legislation areas. It is not intended for real world use.

D. Methodologies

Tags questionnaires aim to provide legal and technical expertise. As such, developing them requires collaboration between legal scholars and security experts. Technical parts tend to have simple implications on the decision graph, mostly deciding between tag values such as encryption requirements and authentication method classes. Legal aspects, on the other hand, requires detailed question flows which may create complicated decision graphs.

Our questionnaire development process so far has been as follows: First, the relevant laws and regulations were studied, and an initial version of the tag space was constructed. Then, the decision graph was created, either using a word processor or a text editor. While the questionnaire was not written using Tags language, it did contain instructions with similar semantics, such as “set” or “go to question 12b”. Once an initial version of the questionnaire was ready, it was translated to Tags, and visualized. The visualization was sent to the legal scholars for vetting. When errors were discovered, the legal scholars sent them back to the coding team. After a few iterations the questionnaire was loaded into a TaggingServer for final QA.

As Tags coding tools advances, we expect this methodology to change. CliRunner, and a language package for the Atom text editor, enable the development of questionnaires directly in Tags. Since Tags programs are plain texts, the development process can leverage distributed version control systems and code review tools for collaboration (see Figure 6).

Decision graph query tools, such as the find-runs command presented in Subsection V-B, enable an new set methodologies, inspired by formal software verification. For example, after studying the relevant jurisprudence and creating the tag space for the questionnaire, the development team can create a formal specification by composing a set of assertions that hold for any valid questionnaire. For example, such assertion might be “no series of answers exists such that the result of the questionnaire will contain the values DataType=personal; Harm=criminal; Transit=clear”. This translates to find-runs DataType=personal; Harm=criminal; Transit=clear returning 0 results.
During the questionnaire development, the team can test it against the set of assertions. This can be seen as a form of test-driven design.

VI. RELATED WORK

Formalization of data handling policies is not new. W3C’s recommendation Platform for Privacy Preferences Project (P3P) [1] allows web servers to state the data they collect and its intended use. It includes a policy matching concept, allowing users to specify which policies they agree to. Another aim of P3P is to allow users understand data handling policies. To this end, version 1.1 [19] (currently a W3C note) adds policy display guidelines. Unlike P3P, which focuses on web browsing scenario, ODRL [2] models policies used by commercial, educational, and government institutes. ODRL policy descriptions are composed of permissions and prohibitions of actions on assets, applying to a user with a defined role. To better support communities with specific needs, the ODRL core is extended by profiles. The PrimeLife Policy Language (PPL) [3], which focuses on downstream data usage, is another policy language that uses rules to permit and deny actions. Its data handling preferences (the way users expect their data to be handled), and data handling policies (the way servers treat the data) are comparable to Tags’ compliance() and support() sub spaces. Data-Purpose Algebra [11] offers a more mathematical approach for modeling data usage restrictions. As ODRL, it aims to model general restrictions, created by legislation, contracts, or “common decency”. DPA describes how data usage restrictions transform when one data item is produced by processing other data items.

The above projects, as well as Tags, aim to prevent forbidden usage. An alternative approach, presented in [17] and [18], is to hold users accountable for such usages.

Seneviratne, Kagal, and Berners-Lee demonstrated a tool for helping users comply with licenses when writing attributions for web content [14]. The design of TaggingServer, which offers a user friendly experience uncommon in data handling policy related tools, was in part inspired by commercial systems like TurboTax, which guide users through filling their tax forms.

VII. FUTURE WORK

DataTags and legal modeling using DHP spaces are very young concepts, and we believe they can be developed in many ways. We are still lacking structured methodologies for developing tag spaces to model legislation and best practices. A study about the limits of this approach is also needed — what are the legal and technological concepts that cannot be captured by a DHP space?

While code tools go a long way in providing a collaborative environment, they still assume a certain technical knowledge, and are not as user friendly as word processors. We envision an on-line collaborative questionnaire development environment, allowing scholars from different backgrounds to develop, discuss, and publish questionnaires, no technical knowledge required. Additional tools can be developed, such as advanced graph queries, automated questionnaire verification, or automated detection of questionable data handling policies.

Finally, Tags’ formal modeling may be useful as a tool for purposes other than dataset handling, such as comparative legal studies.

ACKNOWLEDGMENT

The authors thank Sean Hooley, Alexandra Wood, David O’Brien, Stephen Chong, Sahil Vadhyan, Gary King, and the other members of the Privacy Tools Project at Harvard. We also thank the students of 2016 Fall semester at Ben-Gurion University of the Negev, Israel, who participated in class 202–1–4531 Topics in DataTags. This work was funded by grant CNS-1237235 from the National Science Foundation.

REFERENCES


